Wide-Area Measurement System-Based Low Frequency Oscillation Damping Control Through Reinforcement Learning

Yousuf Hashmy, *Student Member, IEEE*, Zhe Yu, *Member, IEEE*, Di Shi, *Senior Member, IEEE*, and Yang Weng, *Member, IEEE*

***Abstract*—Ensuring the stability of power systems is gaining more attention today than ever before due to the rapid growth of uncertainties in load and increased renewable energy penetration. Lately, wide-area measurement system (WAMS)-based centralized controlling techniques are offering flexibility and more robust control to keep the system stable. WAMS-based controlling techniques, however, face pressing challenges of irregular delays in long-distance communication channels and subsequent responses of equipment to control actions. This paper presents an innovative control strategy for damping down low-frequency oscillations in transmission systems. The method uses a reinforcement learning technique to overcome the challenges of communication delays and other non-linearity in wide-area damping control. It models the traditional problem of oscillation damping control as a novel faster exploration-based deep deterministic policy gradient (DDPG-S). An effective reward function is designed to capture necessary features of oscillations enabling timely damping of such oscillations, even under various kinds of uncertainties. A detailed analysis and a systematically designed numerical validation are presented to prove feasibility, scalability, interpretability, and comparative performance of the modelled low-frequency oscillation damping controller. The benefit of the technique is that stability is ensured even when uncertainties of load and generation are on the rise.**

***Index Terms*—Wide-area networks, low frequency oscillations, damping control, reinforcement learning.**

I. Introduction

AN IMPORTANT concern related to small signal stability is the inter-area oscillations involving various generator groups swinging around each other [1]. Such an effect adversely impacts the economical and reliable operation of an interconnected large-scale power system because the maximum available transfer capability (ATC) is known to be limited [1]–[3]. To alleviate this problem, power engineers use traditional power system stabilizers (PSSs) to try to damp down such oscillations. These inter-area modes, however, are neither always controllable nor observable from local measurement signals [4]. With the development of wide-area measurement systems (WAMS) and practical implementation of phasor measurement units (PMUs), a natural progression to use wide-area damping controller (WADC) through long-distance signal transmission is gaining traction [5]–[7].

Specifically, [8] provides a detailed account for the real-time testing mechanism for POD (phasor-based oscillation damping [9]) algorithm for WADC using OPAL-RT's eMegasim. In [10], we are provided an account of the proposed centralized non-linear controller without considering latency in communication. The work in [11] takes the latency into consideration using a mixed $\frac{H_2}{H_\infty}$ synthesis technique for the design of WADC. Special properties of swing modes derived analytically from a simplified system model are utilized for the control strategy elaborated in [12]. In [13], the authors attempt to address the compensation of delay, as well as the packet dropout issue so a new WADC is proposed. For an overview of WADCs and their operations through different communication channels, we can refer to [14]–[17].

Stochastic WADCs represent another approach to overcome uncertainties. To capture delay uncertainties stochastic control method is adopted by [18], with an unfairly long stochastic delay assumption. To overcome that assumption, [19] uses the expectation modelling of time delays for shorter stochastic time delays. Reference [15] employs a network predictive control scheme for estimating the time delays, which is an extension of [16] as it assumes a low-order model for a more complex power system through least-squares-based identification algorithm. Here, the timing delays are predominantly dependent on the multiple factors, including the power system, as well as those that are not dependent on the system such as difference in communication channels. In [16] timing delay uncertainties are considered in low-frequency oscillation damping control signals, but this assumes the delays as part of system uncertainties and embeds the compensation in the controller. Whereas, the performance of methods in [16] and [20] are highly sensitive to the system operating conditions, that

Manuscript received January 23, 2020; revised May 27, 2020; accepted July 4, 2020. Date of publication July 10, 2020; date of current version October 21, 2020. This work was supported by SGCC Science and Technology Program under Project "AI Based Oscillation Detection and Control" under Contract SGJS0000DKJS1801231. Paper no. TSG-00112-2020. *(Corresponding author: Yang Weng.)*

Yousuf Hashmy is with the AI and System Analytics, GEIRI North America, San Jose, CA 95134 USA, and also with the Department of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: shashmy@asu.edu).

Zhe Yu and Di Shi are with the AI and System Analytics, GEIRI North America, San Jose, CA 95134 USA (e-mail: zhe.yu@geirina.net; di.shi@geirina.net).

Yang Weng is with the Department of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: yang.weng@asu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2020.3008364

TABLE I
DELAYS DUE TO DIFFERENT COMMUNICATION LINKS [17]

Communication Link	One way delay (ms)
Fiber-optic cables	≈ 100-150
Microwave links	≈ 100-150
Power line carrier (PLC)	≈ 150-350
Telephone lines	≈ 200-300
Satellite link	≈ 500-700

greatly threatens the reliability of such a scheme. In addition to the delay uncertainties, renewable energy sources and variations also add stochasticity to the system [21]–[23].

In [24], we are introduced to a variable loop gain controller based on the excessive regeneration for system stability, limiting the delay range up to 250 ms. An approach to identify a low-order transfer function model of a power system using a multi-input multi-output (MIMO) autoregressive moving average exogenous (ARMAX) model is presented in [25]. The use of static VAR compensator (SVC) for damping control are explained in [26]–[28] under different operating conditions and renewable energy sources.

A large-scale power system is a complex, non-linear, and high-order dynamic system, which makes it difficult to obtain the full-scale and detailed model and system parameters. To overcome such issues, the genetic algorithms are implemented to damp down the LFOs [29]. However, this adds to the complexity and unreliability in seeking out global optima without extensive guarantees. Deep supervised learning for WADC is suggested in [30]. Supervised learning methods merely exploit the value of data but do not explore the system, thus lacking adaptability to the previously unknown randomness and transforming nature of the system. Exploration is the key concept in reinforcement learning. Q-learning is proposed to overcome the randomness in the delays for WADC as in [31]. However, Q-learning will most likely be useless when the state space is very large. Larger state spaces are well handled by implementing the gradient directly on the continuous action space.

We propose a controlling agent that can provide a stochastic control using wide-area measurements, so that we can ensure extended observability and flexibility to achieve centralized control. Here, we consider various channels of communication with stochastic delays. Such time delays can be represented by some mixed Gaussain distributions, as shown in Fig. 1. The delay value at any time instant can be anything within the ranges mentioned in Table I for which the probability is non-zero; however, its exact value is highly uncertain. The controller aims to explore the wide-area system and adopts a continuous action to effectively damp the LFOs under different communication delays. This method is similar to deep deterministic policy gradient (DDPG) as in [32]. However, a direct implementation of the method on a large WAMS will take too much time to learn under system uncertainties. Therefore, we provide the solution based on a prior distribution of different time sequences of complete action space to effectively reduce the training computations. We call this implementation DDPG-S. Table II depicts the comparison among different reinforcement learning techniques.

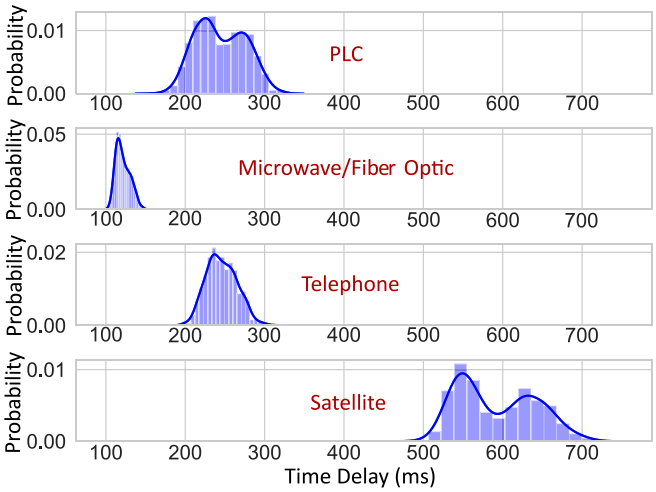

Fig. 1. Probability distributions $P(t_d)$ of delays in different communication links, assuming they are mixed Gaussian distributions.

TABLE II
COMPARISON OF FOUR REINFORCEMENT LEARNING APPROACHES

Item	World model [33]	Q-learning [34]	DDPG [32]	DDPG-S
Model-based (MB) or model-free (MF)?	MB	MF	MF	MF
Need system model and parameters?	Yes	No	No	No
Value function approximation?	No	Yes	No	No
Action Space	Cont./ Discrete	Discrete	Cont.	Cont. Reduced

Note: Here Cont. stands for continuous.

We designed an extensive numerical validation of the proposed technique and tested our learning method on the Kundur's 4-generator system and 10-generator New England system. The performance of the WADC is assessed under different communication delay probability distributions, as shown in Figure 1, and renewable energy penetration to establish the effectiveness of the proposed DDPG-S controller. The impact of different types of faults and locations of the faults are also studied in detail. Furthermore, we present the reduced number of episodes required for effective learning under different communication delays of DDPG-S.

The rest of this paper is organized as follows. Section II gives modelling and problem definition. Section III entails the control scheme. Section IV gives numerical validation of the proposed methodology and Section V concludes the paper.

II. MODELLING

The WAMS-based system provides the ingredients for building up efficient controllers. Phasor measurement units (PMUs) are gaining wide popularity in the transmission systems. The PMUs located at the generator buses detect the voltage and current phasors and estimate the rotor speeds [35]. Here, we assume that each generator-bus is equipped with a PMU so that its phase angle and speed of the generator are available to the controller, with possible delays. Define the state s_t for

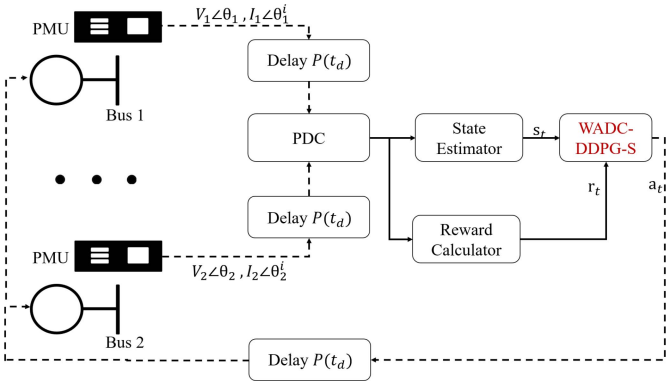

Fig. 2. Framework of the overall scheme. In the above diagram, the dotted lines indicate the communication lines with delays. The inputs to the controller are the states and reward, whereas the output is the action.

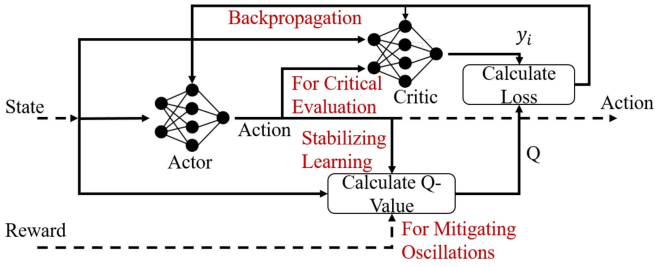

Fig. 3. Zoomed-in illustration of DDPG-S-proposed algorithm. The dotted lines are representative of the inputs and output of the controller.

all observable generators $g = 1, \ldots, G$ to be controlled. The deviations in generator speeds are ω_g^t and the phase angles are θ_t^b between the voltages of the buses $b = 1, \ldots, B$ at remote locations for time $t = 1, \ldots, T$. As the speeds of generators vary upon the occurrence of the disturbance, we use those deviations of the speeds $\Delta\omega_g^t = |\omega_g^t - \omega_g^{t-1}|$ to define the state.

$$\begin{aligned} s_1^t &= \left\{\Delta\omega_1^t, \Delta\omega_2^t, \Delta\omega_3^t, \ldots, \Delta\omega_G^t\right\}, \\ s_2^t &= \left\{\theta_1^t, \theta_2^t, \theta_3^t, \ldots, \theta_B^t\right\}, \\ s_t &= s_1^t \cup s_2^t. \end{aligned} \tag{1}$$

The modern-day PSS is responsible for damping down LFOs by adjusting the voltage applied at the field windings V_g of all the synchronous generators g. As a result, the output of the controller will essentially be an action vector a_t for all the generators g at time t. The action vector a_t, defined in equation (2), is a stabilizing voltage parameter that alters the field voltage of synchronous generators.

$$a_t = \left\{V_1^t, V_2^t, V_3^t, \ldots, V_G^t\right\}. \tag{2}$$

States and actions enable us to completely define the problem.

A. Problem Definition

Problem: Damp down low-frequency inter-area oscillations by adjusting field voltages of generators.

Given:

- a transmission system as the environment X,
- state of the system s_t in X, and communication time delay probability distribution $P(t_d)$.

To Find:

- discounted reward $R(s_t, a_t)$,
- and a policy π comprising of action set a_t for stabilizing field voltages V_g^t of synchronous generators.

B. Identifying Inputs/Output

With the above defined problem, the inputs and outputs for the controller can be established with certainty. The states of the system are obtained directly by manipulating the measurements of voltages and currents of the busses obtained from the PMUs. Moreover, the output of the controller is expected to be a valid control action. The action is comprised of the stabilizing field voltages.

$$\begin{aligned} \text{Inputs} &: \{s_t, r_t\} \\ \text{Output} &: \{a_t\} \end{aligned} \tag{3}$$

where s_t is the state of the system, r_t the reward function which will be designed in the following section, and a_t the control actions.

Fig. 2 gives a complete account of the mechanism, along with the inputs and outputs for the model. The PMUs connected to the remote buses send data (voltage and current phasors) to the phasor data concentrator (PDC), over different communication channels. This data helps in determining state and reward at every time step. Fig. 3 shows the zoomed-in structure of the controller, based on DDPG-S. Furthermore, we design a controller that can produce a high fidelity control action when the states and rewards are fed as input. We aim to design such a controller in the following sections.

III. Control Scheme

Using a model, we aim to develop a reinforcement learning-based robust control scheme. However, this is not possible unless we have a specialized reward design that can maximize the potential of the power system knowledge.

A. Reward Design With Maximized Information Gain

With states, actions, and policy clearly defined, we require an evaluation function that helps in deciding the extent of the fidelity of generated control action. The evaluation function is also called a reward function in the reinforcement learning domain. We design reward that can help maximize the information obtained from the wide-area based observations from synchrophasors and local measurements from the generators. Our goal is to minimize the oscillations in the frequencies of the power system. Based on this, we propose to capture all the features related to oscillations, such as deviation of generator speed from 1 p.u. and abrupt changes in the generator speed with respect to time. However, that will not be enough to capture the effect of buses connected by long-distance lines, so we improve the information gain by leveraging the wide-area measurements of phase angle variation between remote buses as well. We incorporate all this information into the reward design.

Algorithm 1: Novel DDPG-S algorithm for WADC Agent Based on [32]

Randomly initialize critic network $Q(s, a|\theta_Q)$ and actor network $\mu(s|\theta_\mu)$ with weights θ^Q and θ^μ;
Initialize target network with weights $\theta^{Q'}$ and $\theta^{\mu'}$;
Initialize experience buffer $\mathcal{D}$;
for *episode*=1, E **do**
  Receive initial state s_1;
  Initialize a random exploration noise $\mathcal{N}$;
  for $t = 1, T$ **do**
    Set $\mathcal{N}_t^* = \begin{cases} \mathcal{N}_t, & t \in \text{applicable time sequences} \\ 0, & \text{Otherwise} \end{cases}$;
    Select a_t according to the current policy and exploration noise $\mathcal{N}_t^*$;
    Execute action a_t and observe reward r_t and new state s_{t+1};
    Store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$;
    Sample a random minibatch (with size N) of transitions (s_i, a_i, r_i, s_{i+1}) from $\mathcal{D}$;
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$;
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$;
    Update the actor policy using the sampled policy gradient: $\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu_{s_i}} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i}$;
    Update the targets: $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$, $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$;
  end
end

Since the learning agent requires an immensely large action space, which will impede its performance and speed, we further boost the information gain by incorporating the knowledge from locally used power system stabilizers (PSSs) for each generator g. Such information is embedded into the reward in the form of the bounds u and v, where u is indicative of the upper bound of the control action space. Similarly, v represents its lower bound. Such a definition enables the model to treat the action inside and outside the constraints separately.

The speeds of generators are perturbed upon creating the fault in any system, and the low-frequency oscillations should be damped down. For that reason, the reward function consists of four terms. The first terms help to bring the speed ω_g^t as close to 1 pu as possible. The second terms overcome the sustaining deviations in the speeds of the generators. These terms are leverage the information from conventional local PSSs. The third terms in equations (5) and (6) refer to the operation of localized PSS, which helps in limiting the bounds of actions. The fourth terms incorporate the difference between the phase angles of voltages at remote buses. The control actions are chosen based on the oscillations between the remote buses. We use the difference of the phase angles of remote buses to increase the observability because angle differences of remote buses were unobservable without wide-area damping controller. We intend to reduce such a difference so that deviations of speeds of generators connected to remote buses are limited. For all the terms, we take absolute values to capture only the absolute difference so that the highest attainable reward is 0.

As we have established the characteristics that capture the maximum information, we then aim to combine them. A linear relationship among them is the most suitable since we need to have the differences large enough whereby the learning agent can improve by gaining a reasonably large reward. After experiments, the reward values are too small if higher-order terms are employed, and the learning agent will keep oscillating instead of gaining a substantially large reward. So, we summarize the discussion mathematically as,

$$r(s_t, a_t) = \begin{cases} r_1(s_t, a_t) & \text{if } a_t < u, \\ r_2(s_t, a_t) & \text{if } a_t > v, \\ r_3(s_t, a_t) & \text{Otherwise.} \end{cases} \tag{4}$$

$$r_1(s_t, a_t) = \sum_{b=1}^{B}\sum_{t=0}^{T}\sum_{g=1}^{G}\Big(-\alpha\left|1-\omega_g^t\right| - \beta\left|\Delta\omega_g^t\right| - \eta\left|-a_g^t - u\right| - \zeta\left|\theta_b^t - \theta_{b+1}^t\right|\Big), \tag{5}$$

$$r_2(s_t, a_t) = \sum_{b=1}^{B}\sum_{t=0}^{T}\sum_{g=1}^{G}\Big(-\alpha\left|1-\omega_g^t\right| - \beta\left|\Delta\omega_g^t\right| - \eta\left|a_g^t - v\right| - \zeta\left|\theta_b^t - \theta_{b+1}^t\right|\Big), \tag{6}$$

$$r_3(s_t, a_t) = \sum_{b=1}^{B}\sum_{t=0}^{T}\sum_{g=1}^{G}\Big(-\alpha\left|1-\omega_g^t\right| - \beta\left|\Delta\omega_g^t\right| - \zeta\left|\theta_b^t - \theta_{b+1}^t\right|\Big), \tag{7}$$

where α, β, η and ζ are the scaling factors that depend greatly on the system. All of these parameters can be tuned by cross-validating the performance of the agent.

We aim to use future reward values to determine the state and action pairs resulting in a maximum expected reward. As such, we require the rewards of the future and sum them all, but this creates a mathematical difficulty of getting an infinite reward. This condition may happen in the power system control domain very frequently because there exists an infinite number of possible states when we aim to continuously control the generator field voltages. We avoid such a situation by using discounted reward, which gives lower weights to the rewards associated with the state and action pairs in the timesteps of distant future and gives higher weights to the reward values in the near future. Hence, we take the definition of discounted future reward from [36] as R_t

$$R_t = r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \gamma^2 r(s_{t+2}, a_{t+2}) + \ldots, \tag{8}$$

where discount factor $\gamma \in [0, 1]$ is a hyperparameter. The main objective of reinforcement learning is to maximize the expected discounted future reward. Therefore, we need to have a model with such functionality, under time and computational limitations.

B. Learning Agent Achieving Accurate Control

State s_t is involved in defining state value function $V(s_t)$. However, we intend to incorporate more information. For a

learning agent, we propose to use $Q(s_t, a_t)$ so that not only states but also the action values are also considered while making the optimal control decision,

$$Q(s_t, a_t) = E[R_t|s_t, a_t], \tag{9}$$

where $E[.]$ is the expectation function, R_t is the discounted reward, which is already defined in equation (8).

The proposed learning agent is expected to learn the optimal policy π taking the communication delay into account. This is important because there can be uncertainties in the response of generating units and communication channel delays. Such changes become worse with the aging of the equipment. Therefore, we use the learning agent that can learn through interacting with the power system and eventually learns to deliver the policy based on maximized information. Unlike a conventional feedback control system designed to respond in a certain fashion based on the current state of the system, we propose to embed the knowledge from the past and random exploration of the transmission network so that the uncertainties due to load switching, capacitor bank switching, and DERs can be learned effectively. Moreover, the adaptation of the learning agent to the ever-dynamic nature of the power system is also significant due to the dynamics of power systems.

To accomplish the goals, the discretized states and actions will help to generate a discrete Q function, but in a real scenario such as a power system, the number of states and action pairs can grow rapidly with an increase in the count of buses in the system. This may create a Q table of an infinite number of entries. To overcome this challenge, we introduce a function approximation method, based on neural networks to provide a finite set of parameters that can be learned through experience data. The data comprises of the states and actions of the current and future states along with their respective reward values. Then, we are in a better position to approximate the Q function.

However, using only a single neural network might fail because there will be a high chance of falling victim to local minima. Therefore, as shown in Fig. 3 we choose an actor-critic model, where a critic-network might help to suppress the bad decisions made by an actor-network. The details of continuous control through the actor-critic model is presented in [32]. Our proposed model not only deals with continuous action space, but also its fidelity for learning under complex environments is proven in [36]. By combining expressions (9) and (8), we obtain

$$Q(s_t, a_t) = E[r(s_t, a_t) + \gamma \max_a E\big[Q(s_{t+1}, a_{t+1})\big]. \tag{10}$$

The function approximator Q for the critic network, by sampling states from the wide-area measurements following a specific distribution is given in [32]. An actor-network $\mu(s_{t+1})$ takes only the states as input features and directly estimates the actions. But such estimation requires critical evaluation. So, we define the approximator y_i for critic network,

$$y_i = r_i + \gamma Q'\Big(s_{i+1}, \mu'\Big(s_{i+1}|\theta^{\mu'}\Big)|\theta^{Q'}\Big) \tag{11}$$

Since both of the function approximators are characterized as deep layers of neural networks, we parameterize them with

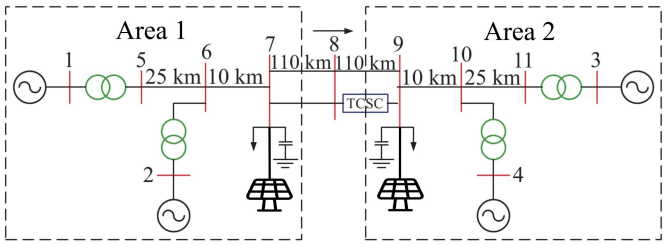

Fig. 4. Two-area (4-generator) system from [37] with additional renewable energy integration.

θ^Q and θ^μ as presented in [32]. So, the loss function for M samples becomes,

$$\text{Loss} = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(s_i, a_i))^2, \tag{12}$$

where $i = 1, \ldots, M$ is the number of samples in mini-batch. The parameters of critic-network are obtained by iteratively minimizing the above loss function [32]. The goal in reinforcement learning is to learn a policy that maximizes the expected return from the start distribution J. We update the actor-network by applying the chain rule to the expected return from J concerning the actor parameters,

$$\nabla_{\theta_\mu} J \approx \frac{1}{M}\nabla_a Q\Big(s_i, a_i|\theta^Q\Big)\nabla_{\theta_\mu}\mu\big(s_i|\theta^\mu\big), \tag{13}$$

Then, we update the target actor Q' and target critic μ' parameters using a periodic approach, so that after each iteration the target actor becomes the initial actor and target critic becomes the initial critic as proposed in [32], $\theta^{Q'} = \theta^Q$ and $\theta^{\mu'} = \theta^\mu$ respectively.

The flow of the algorithm for damping the low-frequency oscillations using this procedure can also be observed in Fig. 3. Algorithm 1 provides a detailed account of the working mechanism of the proposed methodology for effective control policy learning in a reasonable amount of time. The optimal values of hyperparameters including discounting factor, experience buffer size and minibatch size are determined by extensive cross-validation of the model. The actions a_t reach the generators after the time delays taken from the probability distribution $P(t_d)$ for each kind of communication delay. The non-linearity is an unknown for the model to train on. A fully trained model is capable of adjusting itself to the delays in communication.

C. Theoretical Proof on Stability Guarantee

There has been an ongoing discussion on the stability guarantee of a controller relying on the neural networks and reinforcement learning. Machine learning algorithms are dependent upon the data. Specifically, reinforcement learning algorithms gather data by exploring the environment directly. Hence, the environments are the key factor in establishing the guarantee of stabilization. The larger the available data is and the more realistic the environment is, the better is the performance of the reinforcement learning systems.

We let a DDPG-S powered controller ρ, having states $s_t \in \mathbb{R}^{(G+B)\times T}$ and $a_t \in \mathbb{R}^{G\times T}$. By assuming an exploration

e summed up with policy π parameterized by θ^Q and θ^μ, we can get the specific state of the environment.

$$u_t = \pi\left(y_t|\theta^Q, \theta^\mu\right) + e_t, \tag{14}$$

where e_t represents the exploration that captures additive randomization effect, y_t and u_t depicts the input and output respectively. The objective of RL is to maximize expectation of reward r_t. We can simply assume the energy of e_t to be bounded over time $\|e\|_2 = \sqrt{\int |e_t|_2^2 dt} \leq \infty$. Since ρ deals with implementing gradient directly on π and is a neural network based learner, the stability criteria is expressed in terms of L_2 gain [38].

Definition 1: The L_2 gain of the environment X controlled by π is the worst-case ratio between total output energy and total input energy:

$$K(X, \pi) = \sup_{s \in L_2} \frac{\|y\|_2}{\|u\|_2}, \tag{15}$$

where L_2 gives all the square-summable signals. The total energy over time is $|y|_2 = \sqrt{\|y_t\|_2^2 dt}$, and equation (14) gives the state and learner relationship with exploration. Whenever $K(X, \pi)$ is finite, the interconnected system is said to have input-output stability (or finite L_2 gain) [38].

$$\rho\{\epsilon\} = \left\{\pi \,|\overline{\epsilon}_t \leq \partial\pi_t \leq \underline{\epsilon}_t\right\} \tag{16}$$

as the controller whose policy has its partial derivative bounded by $\overline{\epsilon}_t \in \mathbb{R}^{(G+B)T \times GT}$ and $\underline{\epsilon}_t \in \mathbb{R}^{(G+B)T \times GT}$, it is desirable to provide stability certificate as long as the RL policy remains within the above "safety set." Specifically, consider the linear time invariant component K:

$$\begin{aligned} \dot{x}_K &= Ax_K + Bs + v, \\ y &= x_K, \end{aligned} \tag{17}$$

where $x_K \in \mathbb{R}^{(G+B)T}$ is the state and output is $y \in \mathbb{R}^{(G+B)T}$. Since system is assumed to be stable, A is Hurwitz.

$$\begin{aligned} s &= e + a, \\ w &= \pi(y), \\ v &= p_t(y), \end{aligned} \tag{18}$$

where $e \in \mathbb{R}^{G \times T}$ is the exploration, $a \in \mathbb{R}^{G \times T}$ is the policy and $p_t : \mathbb{R}^{(G+B)T} \to \mathbb{R}^{(G+B)T}$ represents the uncertainty added into the system due to communication delays having respective probability distributions $P(t_d)$. It is assumed to satisfy integrated quadratic constraint (IQC) and represented by (Ψ, M_{p_t}) [39], and Ψ has the state space of,

$$\begin{aligned} \dot{\psi} &= A_\psi \psi + B_\psi^v v + B_\psi^y y, \\ z &= C_\psi \psi + D_\psi^v v + D_\psi^y y, \end{aligned} \tag{19}$$

where internal state is $\psi \in \mathbb{R}^{(G+B) \times T}$ and filtered out state is z. By defining $x = [x_K^T \;\; \psi^T]^T \in \mathbb{R}^{2*(G+B) \times T}$ as the newly created state, and assuming $w = W_q$.

Theorem 1: Let transmission network X be stable (i.e., A is Hurwitz) and $\pi \in \mathbb{R}^{(G+B)T \times GT}$ bounded causal controller. Assume that:

1) the interconnection of transmission network X and controller policy π is well-posed;
2) π has bounded partial derivatives on the open subset $\mathbb{B}$ (i.e.,$\overline{\epsilon}_t \leq \partial\pi_t \leq \underline{\epsilon}_t, \forall x \in \mathbb{B}$).
3) $p_t \in \mathrm{IQC}(\Psi, M_{p_t})$, where Ψ is stable.

If semi-definite programming SDP($\mathbb{P}, \lambda, \Gamma, \epsilon$) is feasible for $\mathbb{P} \geq 0$ and $\Gamma > 0$, then the feedback interconnection of the nonlinear system and policy π is stable upon satisfying stability condition from [38],

$$\int_0^T |y_t|^2 dt \leq \Gamma^2 \int_0^T |e(t)|^2 dt. \tag{20}$$

Proof: Using newly created state Ψ and the filtered output z as the constraints of IQC on delay uncertainties p_t following [39]. The exact solution of SDP($\mathbb{P}, \lambda, \Gamma, \epsilon$) is feasible, and is available in [38]. We obtain dissipation inequality by multiplying both sides of SDP($\mathbb{P}, \lambda, \Gamma, \epsilon$) with $[x^T \; q^T \; v^T \; e^T]^T$ and $[x^T \; q^T \; v^T \; e^T]$.

$$\frac{dx_K^T \mathbb{P} x_K}{dt} + z^T M_{P_t} z + \begin{bmatrix} x_G \\ p_t \end{bmatrix}^T M_\pi \begin{bmatrix} x_G \\ p_t \end{bmatrix} < \Gamma e^T e - \frac{1}{\Gamma} y^T y \tag{21}$$

Because $p_t \in \mathrm{IQC}(\Psi, M_{p_t})$, $z^T M_{P_t} z$ is non-negative. The third term on the left-hand side of the inequality is guaranteed to be non-negative due to the property of smoothness quadratic constraint [38],

$$\frac{dx_K^T \mathbb{P} x_k}{dt} - \Gamma e^T e + \frac{1}{\Gamma} y^T y < 0. \tag{22}$$

Since time derivative of the storage function is negative definite, first term on left is guaranteed to be negative.

$$\frac{1}{\Gamma} y^T y < \Gamma e^T e. \tag{23}$$

By integrating both sides from 0 to T time of each episode, we see that inequality (20) holds. Hence, the theorem is proven. ■

In [38], there is a comprehensive account of the safe limits to ensure performance guarantee. Additionally, in [38], a preventative certificate of stability for a broad class of neural network controllers including policy gradient-based algorithms is presented in detail.

D. Action Space Reduction for Improving Computation Time

The learning model has to be accurate, but its training speed is also important. By considering time domain, we get an immense action space. To have a working algorithm, we need to remove most of the infeasible action space in the exploration phase, using our domain knowledge, to ensure high performance in lesser time. A conventional DDPG algorithm usually provides solution for the problems, which are mostly games that do not contain the time dimension. However, in the damping control issue where time domain is highly significant, we introduce the time dimension to embed the physical law into the algorithm – eliminating the physically infeasible region and enhancing the exploitation in the physically feasible region. Therefore, instead of directly utilizing the standard action selection method $a = \mu(s) + \mathcal{N}$, where $\mathcal{N}$ is stochastic noise from the noise model,we restrict it based on time points or sequences. For example, in our domain, the amplitude of oscillations is initially large, while it recedes with time. Hence,

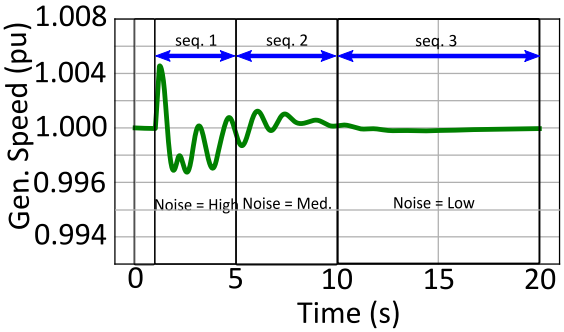

Fig. 5. Time sequences-based division of stochastic Noise $\mathcal{N}_t$ reducing the action space. Here seq. stands for time sequences and Med. stands for medium.

we use a specific distribution of noise divided based on time sequences.

$$\mathcal{N}_t^* = \begin{cases} \mathcal{N}_t, & t \in \text{applicable time sequences} \\ 0, & \text{Otherwise.} \end{cases} \tag{24}$$

Fig. 5 shows segregation of the time dimension on the basis of time sequences, and employing a different noise for each one reduces the action space while exploring. For time sequence 2 and time sequence 3, we achieve a very significant reduction in learning episodes because the agent finds the optimal policy by exploring in the reduced region. Anything outside these applicable time sequences do not have any noise, and that is equivalent to the time before the occurrence of the fault. We provide a study based on the speed measurement for learning using the mean overall speed of the agent towards the target, which is defined using the success rate. The success rate is the rate at which the episodes end without losing the synchronism.

$$\bar{v} = \frac{\sum v_x cos\left(\theta_{\text{target}}\right)}{v_x^{\max}}, \tag{25}$$

where $\bar{v}$ is the average speed and v_x is the speed value for all the time steps in an episode. The maximum value is named as v_x^{max}. θ_{target} gives the parameters of the target in policy gradient method.

We have the blueprint for the learning agent to sample out the reliable control actions, but this is not only in the conventional working environments, but also where there are irregularities of the renewable energy supply. So we have to find a way to prove the controller is working for damping down the oscillations in time, and in an environment that is prone to different uncertainties such as renewable energy sources. With the special feature of the learning agent, uncertainties in the system can also be compensated accordingly. The unique design of the learning agent enables it to learn how to perform well in a power system setting, that is highly dynamic.

IV. Numerical Validation

In this section, the proposed method was applied to the 2-area and 4-generator Kundur system and the IEEE 39-bus 10-generator system. A PID controller and a networked predictive controller (NPC) [15] are employed as benchmarks.

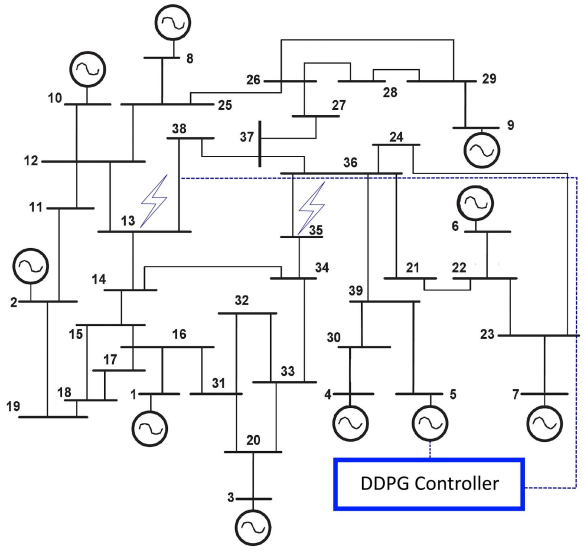

Fig. 6. 39-bus system with controller for the generator at bus 5 by drawing measurements from bus 13.

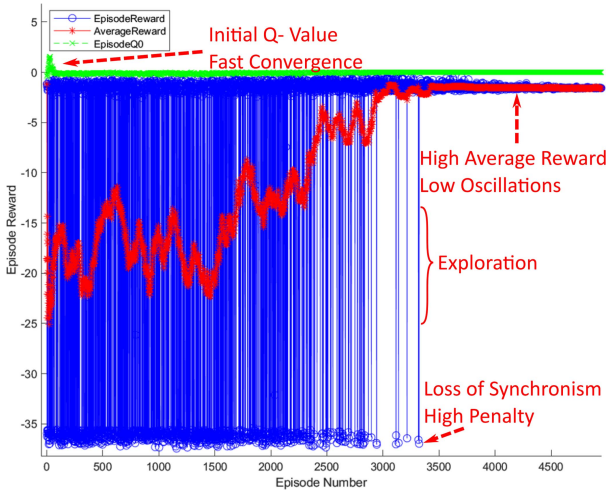

Fig. 7. Learning curve showing reward with respect to increasing episodes. The blue points show the episode reward. Red line is showing the average reward considering an averaging window of past five episodes. Green points indicate the initial Q value for each episode.

A. Validation Setup

For the small-signal studies, the 4-generator model has been used extensively in the past to provide evidence for the methodologies adopted in transmission systems [37] and [40]. The system under study consists of two symmetrical areas linked together by two 230 kV lines of 220 km long. It was specifically designed to study low-frequency electromechanical oscillations in large interconnected power systems. The detailed model is shown in Fig. 4. We control all the 4-generators through our proposed methodology because of the limited computational capability allowed for a 4-generator control.

Despite its small size, the behavior of typical systems in actual operations resemble this test case to a reasonable degree. Each area is equipped with two identical generating units rated 20kV/900MVA [37] and [40]. The synchronous machines have identical parameters, except for inertias, which are $H = 6.5s$

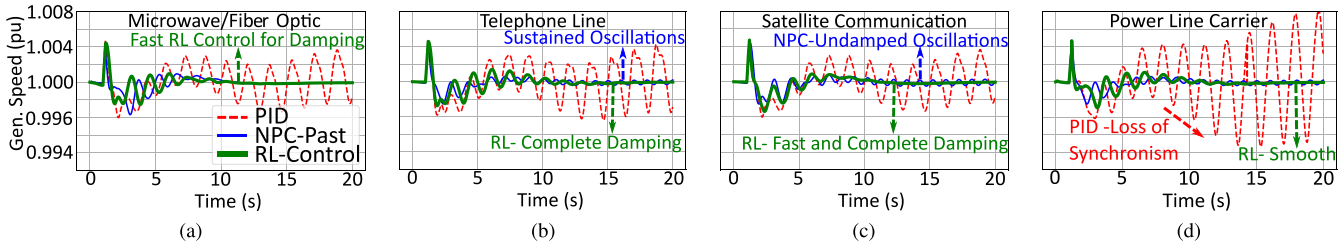

Fig. 8. Control method with reinforcement learning compared with past techniques with only the communication delay uncertainties corresponding to different communication channels.

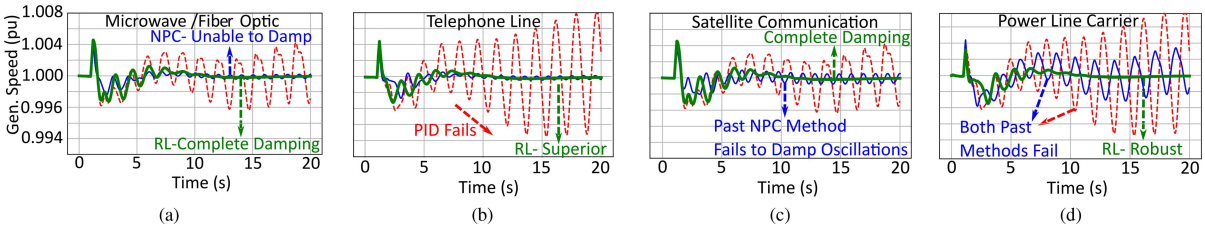

Fig. 9. Control method with reinforcement learning compared with past techniques when renewable energy uncertainties are introduced to the system along with the communication delay uncertainties corresponding to different communication channels.

in the first area and $H = 6.175s$ in the second area. Thermal plants having identical speed regulators are further assumed at all locations, alongside fast static exciters with a 200 gain. The load is represented as constant impedances and split between the areas in such a way that area 1 is exporting 413 MW to area 2. Since the surge impedance loading of a single line is about 140 MW, the system is somewhat stressed, even in steady-state. The reference load-flow, with area 2 considered as the slack machine, is such that all generators are producing about 700 MW each. Additionally, a solar power system of 100 MW is added to both the areas. This creates an uncertain loads variation.

The 10-generator and 39-bus system from New England is also used to validate the proposed methodology. Such a large system helps to establish the scalability of the mechanism. For the disturbance, we select the buses to apply faults that can give rise to oscillations. Since faults on any of the buses will have a similar effect on the controlled generators, as long as at least one long-distance line exists in between the faulty bus and controlled generator. Keeping these points into consideration, we have the liberty to select bus 13 and bus 35, both of which have long-distance lines between them and a controlled generator bus 5. Similarly, the measurements are taken from bus 13 as the input to the reinforcement learning-based controller. The generator under reinforcement learning control is connected to bus 5, as illustrated in Fig. 6. The control mechanism is validated by controlling only one generator while the others are assumed to be controlled through standard PSSs. The reason for controlling only one generator is that the burden on computational capability becomes too high when action space becomes 10 dimensional. The single generator control strategy is widely adopted for LFO damping control, as in [41].

B. Training and Testing

For the training of WADC, we interface the RL-agent block with power system simulation in Simulink. For each episode, the simulation is repeated from 0s to 20s. The deep layers in the neural networks are updated with their optimal weights according to the loss being provided in equation (12). That loss is calculated using a dataset stored in experience buffer $\mathcal{D}$ as depicted in Algorithm 1. The dataset comprises of the features $\{s_t, a_t, r_t, s_{t+1}\}$, with t being a timestep. Initially, the experience is filled with the data, and N random samples are taken for predicting value function. Then real and predicted ones helps to update the critic parameters. A sampled policy gradient updates the actor parameters. The policy π adopted in each episode is evaluated by the reward calculator. The higher the cumulative reward of an episode is, the more the future policies are close to the policy resulting in high reward. If an episode encounters loss of synchronism of the system, the episode terminates immediately with a very high penalty instead of a reward. The training process stops when the difference between episode reward and the average reward becomes less than 0.1. The model is trained under different system operating conditions by simulating for each one, saving the model parameters, and then resuming the training by initializing with previous parameters.

Next, we test out learned agents. The action value corresponding to the specific states of the system are already determined for each time step. The set of optimal action values is called the policy. The policy is reproduced whenever the model is tested under the similar state, resulting in high rewards and least oscillations.

C. Validation: Achieve Accurate Control Under Different Scenarios

With a simulation environment in place, we interface the policy gradient-based reinforcement learning agent to the measurements from the state evaluator. The control action is supplied to the field voltage of the generators under consideration. The four generator model results are shown in Fig. 9, due to space limitations.

Equations (4)-(7) show that the maximum attainable reward is 0. We applied the algorithm to a 4-generator system. The learning curve seen in Fig. 7 shows that the model starts at a very low value and then upon learning on based on the discounted reward, the parameters of the neural networks are updated and the policy tends toward optimal value. After $5,000$ episodes, the average reward in Fig. 7 reaches a value close to 0. Since the reward functions are highly dependent on the existence of the oscillations, we aim to show the performance of a well-learned model where there are low-frequency oscillations, in case of different communication channel delays. Additionally, Fig. 7 validates the special reward design we proposed. The high penalty is enforced in the scenarios where the system loses synchronism since such a case will be responsible for a large outage of the system. We consider such scenarios as game-over for the model, and there is no further evaluation performed so that the training time can be curtailed.

The model explores from episode 0 to around episode $2,000$ as illustrated in Fig. 7. Reasonable exploration is the primary requirement while learning in uncertain environments. The model achieves a high average reward after sufficient exploration. After $4,000$ episodes, the model converges to a very high reward. The effect of learning can also be observed from another perspective of implementing a control policy that maintains system synchronization for stable operation of the whole power system. Fig. 7 shows that after $3,400$ episodes, high fidelity control action enables the system to maintain stability.

Another set of experiments are conducted to study the effect of different operating conditions and the efficiency of the learning model. For this reason, we vary the length of the line between two different areas in the 4-generator model. The length of the line is varied at a constant step size of 20 km. Table III illustrates the reward accumulation after $5,000$ episodes since gaining a high reward is the primary job of the learner. A high mean reward of -6.14 and a maximum reward of -0.11 ensure the least oscillations when the line length is 140 km. Furthermore, with the same system parameters and training settings, we develop Table IV, which presents the damping ratios under the NPC (past) method and the proposed RL agent. Here, we vary the length of the line between the two areas of the 4-generator model. The analysis shows that under the influence of PLC delay, there are some instances where NPC and RL (DDPG-S) both perform equally well, e.g., at line length equal to 160 km. For most of the cases, the damping is more abrupt in the case of the learned RL agent.

Since we are interested in evaluating model performance under different fault types, we designed an experiment on the 10-generator model by applying different types of faults on bus 13 and bus 35. The training is carried out until the difference between average and episode rewards is less than 0.1. The accumulated reward values for each episode are recorded. Their means, maximums, and standard deviations (Std.) are determined and tabulated in Table V, which shows some interesting points. For bus 13, the impact on the highest mean reward of -18.59 with a least standard deviation of 41.34 is obtained when the model is tested under the single line to ground (LG) fault as compared to double line (LL) faults, double line to ground (LG) faults, and tripe line faults (LLL). This shows the mildness of the LG case. Moreover, we observe that for bus 35, overall means, and maximums are lower, and standard deviations are relatively high, ignoring very few cases (due to the stochasticity in the system). The reason for such values for bus 35 is the closeness to the generator under control.

TABLE III
EFFECT OF CHANGE OF LINE LENGTH BETWEEN 2-AREAS ON REWARD OF THE AGENT OF 4-GENERATOR MODEL WITH PLC DELAY

Line Length (km)	Mean Reward	Maximum Reward	Std. Reward
220	-10.12	-0.46	36.27
200	-13.08	-0.33	32.40
180	-8.53	-0.37	29.07
160	-9.32	-1.32	29.51
140	-6.14	-0.11	24.93

Note: All the cases are generated by fixing the number of learning episodes to 5,000. The highlighted row shows the results for default operating conditions of 4-generator case.

TABLE IV
COMPARISON OF DAMPING RATIOS UNDER THE LEARNED PL (DDPG-S) AGENT AND NPC METHOD FOR 4-GENERATOR MODEL WITH PLC DELAY

Line Length (km)	Damping Ratio	
	NPC	RL
220	0.025	0.44
200	0.29	0.44
180	0.39	0.45
160	0.46	0.46
140	0.49	0.52

TABLE V
EFFECT OF FAULT TYPE ON REWARD OF THE AGENT OF 10-GENERATOR MODEL WITH PLC DELAY

		Fault Type			
	Reward	LG	LL	LLG	LLL
bus 13	Mean	-18.59	-21.44	-21.88	-19.27
	Maximum	-0.87	-0.59	-0.65	-0.88
	Std.	41.34	48.22	51.73	50.26
bus 35	Mean	-21.39	-21.41	-24.70	-24.27
	Maximum	-0.92	-0.99	-1.35	-1.44
	Std.	52.17	51.62	51.28	53.77

D. Validation: Evaluate Performance Under Communication Delays

With a well-learned model, our next step is to validate the performance of the model with communication delays. In Fig. 8(a) the performance of different algorithms with communication time delay in microwave link or fiber-optic line is shown. The comparison is carried out among an optimized PID controller, a past method based on a network predictive approach (NPC), and the proposed reinforcement learning (RL)-based approach. It is safe to say that the RL Control method has outperformed others. Although we prove the effectiveness of our model under one kind of communication channel, it is imperative to establish the effect of

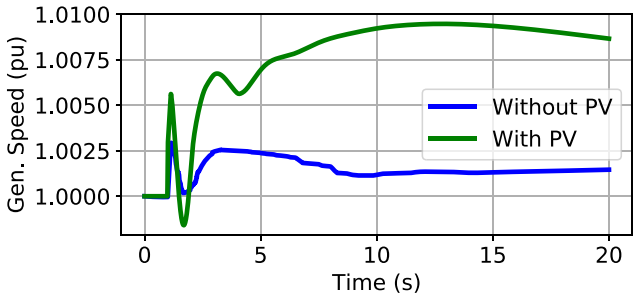

Fig. 10. Difference between the generator speeds with (80%) and without PV penetration.

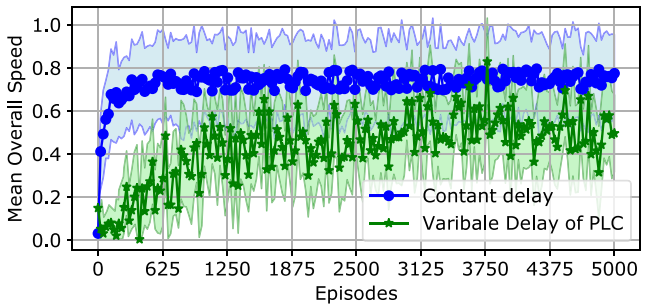

Fig. 11. Mean overall speed $\bar{v}$ of the learning agent with constant time and variable time delay of PLC.

other communication channels available for SCADA measurements. Hence, Fig. 8(b) shows the comparison of the control schemes when the communication delays of telephone lines are simulated. Moreover, Fig. 8(c) and Fig. 8(d) show the performance of the proposed model for a satellite link and PLC, respectively.

PLC has an estimated communication delay between 150 ms to 300 ms. Similar to the other channels, in this case, the proposed controller has the best performance. Moreover, the result in Fig. 8(d) shows that the other methods have shown worse performance, while the proposed method has successfully contained the oscillations in the case of satellite communication delay. This shows that not only has the model achieved a high fidelity in controlling the generator, but also the timing delay randomness is successfully learned. Other methods, such as a tuned PID controller, completely fails when there are significant delays in the communication channels, and the NPC method also suffers deterioration in performance. Hence, to accomplish a wide-area-based controller, deep deterministic policy gradient provides the best solution under normal circumstances.

E. Validation: Assess Efficacy Under PV Uncertainties

In this section, we aim to provide validation of our proposed technique under the presence of up to 80% of load shared by the solar power system in both the areas. The impact of PV in damping control is overwhelming. Fig. 10 provides an idea regarding the impact of PV uncertainties; we plot the conventional PSS control of the 4-generator system with 80% solar penetration.

We retrained the model, with time delay uncertainties and PV sources incorporated into the system. Fig. 9 shows the frequency oscillations by different controllers under the influence of a microwave or fiber-optic link, a telephone line, a satellite link, and power line carrier delays, under the influence of renewable energy uncertainties. The results show that although the increase in uncertainty has made the control more difficult for the PID controller and for past methods, it does not affect the RL control much because the RL method can learn by exploring the environment.

Specifically, for both telephone line and power line carrier communication channels, the tuned PID controller fails to provide a required control. Such a failure causes the speeds of the generators to deviate from the standard 1 pu, and eventually causes the whole system to lose synchronism. For the remaining two channels of microwave or fiber-optic links and satellite links, the deviation of speeds is slower, but the result will be the same, i.e., the loss of synchronism.

Note that the networked predictive control (NPC) shows a better performance than a simple PID controller. However, with more uncertainties introduced into the system, the performance of NPC deteriorates, as indicated in Fig. 9. The communication delays of microwave and fiber-optic, which are well handled in the case without renewable energy integration, are not the case in the presence of renewable energy. The oscillations are not completely damped down in the case of microwave and fiber-optic link, as shown in Fig. 9(a).

The overall performance is not much different, under the delays due to other communication modes. A similar effect can be observed by comparing the results for telephone line, satellite communication, and power line carrier modes. For both PID and NPC controllers, the low-frequency oscillations are not getting damped completely, even after 20 s for the telephone line and satellite communication. The case of power line carrier is even worse, where the oscillations keep growing in amplitude and will result in the loss of synchronism for the whole system. Hence, in the proposed method, the exploration-exploitation learning process with solar power systems can help the model to experience such scenarios as well, and it can adjust the parameters to meet the requirement of damping down the oscillations.

The proposed method based on reinforcement learning has proven its reliability in successfully damping down the LFOs for all kinds of communication channels, even under the presence of uncertainties due to deep penetration of solar power systems. For microwave and fiber-optic link as well as telephone lines, the oscillations are damped down very quickly, whereas, the satellite and power line carriers have relatively slower damping. This proves that DDPG-S has the capability of learning the uncertainties effectively, including not only those that are introduced due to communication delays, but also the ones introduced by renewable energy integration into the system.

F. Validation: Accomplish High-Speed Learning With Uncertainties

The accuracy of the control methodology is not the only consideration while selecting the method, but researchers are also concerned with the speed and computational feasibility of the learning model. Therefore, we analyze the time required

TABLE VI
DOMAIN KNOWLEDGE UTILIZATION FOR ACTION SPACE REDUCTION IN 4-GENERATOR CASE

Comparison	# of episodes taken to reach $\bar{v}$ greater than 0.6			
	PLC	Microwave	Telephone	Satellite
Without Removal	1325	1083	1265	1455
With Removal	412	284	557	817

for learning for the 4-generator system when there is a constant time delay versus a variable uncertain time delay of the Gaussian mixture model. Fig. 11 indicates, that more episodes are required to train the model when the communication delay has uncertainty. Therefore, we provide a model that avails the domain knowledge of the oscillation damping problem in power systems. The overshoots of such oscillations are larger in the first 10 s; we use this time for exploring larger action space. Such a procedure shows encouraging results as seen in Table VI. By removing the unnecessary action space, the number of episodes required to reach the mean overall speed of greater than 0.6 is increased for all the communication channels. This is intuitive because the continuous action space is limited to only the relevant space minimizing the exploration computations.

V. CONCLUSION

Inter-area oscillations have been a longstanding issue in power systems, and controlling them is a challenge that needs to be addressed using modern techniques. The wide-area measurement systems provide a centralized control philosophy; however, it faces serious concerns of the uncertainties in the communication delays and responses time of the equipment. We provide a holistic solution to such a problem by carefully modelling the control methodology and leveraging the capability of learning stochastic continuous control actions through a policy gradient method. Such a policy is learned by employing deep neural networks as the approximator. We provide the discussion on the stability of the learning-based controller. Additionally, the training speed is substantially increased by employing the prior knowledge of time sequences. The proposed methodologyy is validated numerically on numerous test cases under different scenarios of renewable energy penetration. The results prove scalability and robustness of the control system for low-frequency oscillations. Hence, a stable power system is ensured.

REFERENCES

[1] I. Zenelis and X. Wang, "Wide-area damping control for interarea oscillations in power grids based on PMU measurements," *IEEE Control Syst. Lett.*, vol. 2, no. 4, pp. 719–724, Oct. 2018.

[2] S. P. Azad, R. Iravani, and J. E. Tate, "Damping inter-area oscillations based on a model predictive control (MPC) HVDC supplementary controller," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3174–3183, Aug. 2013.

[3] M. Klein, G. J. Rogers, and P. Kundur, "A fundamental study of inter-area oscillations in power systems," *IEEE Trans. Power Syst.*, vol. 6, no. 3, pp. 914–921, Aug. 1991.

[4] M. E. Aboul-Ela, A. Sallam, J. D. McCalley, and A. Fouad, "Damping controller design for power system oscillations using global signals," *IEEE Trans. Power Syst.*, vol. 11, no. 2, pp. 767–773, May 1996.

[5] S. Zhang and V. Vittal, "Design of wide-area power system damping controllers resilient to communication failures," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4292–4300, Nov. 2013.

[6] M. Zahid, Y. Li, J. Chen, J. Zuo, and A. Waqar, "Inter-area oscillation damping and voltage regulation by using UPFC for 500 kV transmission network," in *Proc. IEEE 2nd Int. Conf. Control Robot. Eng.*, Bangkok, Thailand, 2017, pp. 165–169.

[7] I. Kamwa, R. Grondin, and Y. Hebert, "Wide-area measurement based stabilizing control of large power systems-a decentralized/hierarchical approach," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 136–153, Feb. 2001.

[8] E. Rebello, L. Vanfretti, and M. S. Almas, "Experimental testing of a real-time implementation of a PMU-based wide-area damping control system," *IEEE Access*, vol. 8, pp. 25800–25810, 2020.

[9] L. Angquist and C. Gama, "Damping algorithm based on phasor estimation," in *Proc. IEEE Power Eng. Soc. Winter Meeting Conf.*, vol. 3. Columbus, OH, USA, 2001, pp. 1160–1165.

[10] G. Sánchez-Ayala, V. Centeno, and J. Thorp, "Gain scheduling with classification trees for robust centralized control of psss," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 1933–1942, May 2016.

[11] M. Beiraghi and A. M. Ranjbar, "Adaptive delay compensator for the robust wide-area damping controller design," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4966–4976, Nov. 2016.

[12] V. Pradhan, A. M. Kulkarni, and S. A. Khaparde, "A model-free approach for emergency damping control using wide area measurements," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4902–4912, Sep. 2018.

[13] M. Li and Y. Chen, "A wide-area dynamic damping controller based on robust h_∞ control for wide-area power systems with random delay and packet dropout," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 4026–4037, Jul. 2018.

[14] A. Chakrabortty, "Wide-area damping control of large power systems using a model reference approach," in *Proc. 50th IEEE Conf. Decis. Control Eur. Control Conf.*, Orlando, FL, USA, 2011, pp. 2189–2194.

[15] W. Yao, L. Jiang, J. Wen, Q. Wu, and S. Cheng, "Wide-area damping controller for power system interarea oscillations: A networked predictive control approach," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 1, pp. 27–36, Jan. 2015.

[16] H. Wu, K. S. Tsakalis, and G. T. Heydt, "Evaluation of time delay effects to wide-area power system stabilizer design," *IEEE Trans. Power Syst.*, vol. 19, no. 4, pp. 1935–1941, Nov. 2004.

[17] M. R. Younis and R. Iravani, "Wide-area damping control for inter-area oscillations: A comprehensive review," in *Proc. IEEE Elect. Power Energy Conf.*, Halifax, NS, Canada, 2013, pp. 1–6.

[18] X. Zhang, C. Lu, and Y. Han, "Stability analysis of wide-area damping control system with stochastic communication time delay," in *Proc. IEEE Power Energy Soc. Innovat. Smart Grid Technol. Conf.*, Washington, DC, USA, 2015, pp. 1–5.

[19] C. Lu, X. Zhang, X. Wang, and Y. Han, "Mathematical expectation modeling of wide-area controlled power systems with stochastic time delay," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1511–1519, May 2015.

[20] B. Yang and Y. Sun, "Damping factor based delay margin for wide area signals in power system damping control," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3501–3502, Aug. 2013.

[21] X. Zhang, C. Lu, S. Liu, and X. Wang, "A review on wide-area damping control to restrain inter-area low frequency oscillation for large-scale power systems with increasing renewable generation," *Renew. Sustain. Energy Rev.*, vol. 57, pp. 45–58, May 2016.

[22] M. Tajdinian, M. Allahbakhshi, M. Mohammadpourfard, B. Mohammadi, Y. Weng, and Z. Dong, "Probabilistic framework for transient stability contingency ranking of power grids with active distribution networks: Application in post disturbance security assessment," *IET Gener. Transm. Distrib.*, vol. 14, no. 5, pp. 719–727, Mar. 2020.

[23] Y. Weng, Y. Liao, and R. Rajagopal, "Distributed energy resources topology identification via graphical modeling," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2682–2694, Jul. 2017.

[24] D. Roberson and J. F. O'Brien, "Variable loop gain using excessive regeneration detection for a delayed wide-area control system," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6623–6632, Nov. 2018.

[25] H. Liu *et al.*, "ARMAX-based transfer function model identification using wide-area measurement for adaptive and coordinated damping control," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1105–1115, May 2017.

[26] A. Vahidnia, G. Ledwich, and E. W. Palmer, "Transient stability improvement through wide-area controlled SVCs," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 3082–3089, Jul. 2016.

[27] X. Y. Bian, Y. Geng, K. L. Lo, Y. Fu, and Q. B. Zhou, "Coordination of PSSs and SVC damping controller to improve probabilistic small-signal stability of power system with wind farm integration," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 2371–2382, May 2016.
[28] K. Zhang, Z. Shi, Y. Huang, C. Qiu, and S. Yang, "SVC damping controller design based on novel modified fruit fly optimisation algorithm," *IET Renew. Power Gener.*, vol. 12, no. 1, pp. 90–97, Jan. 2018.
[29] M. E. C. Bento, D. Dotta, R. Kuiava, and R. A. Ramos, "A procedure to design fault-tolerant wide-area damping controllers," *IEEE Access*, vol. 6, pp. 23383–23405, 2018.
[30] S. Jhang, H. Lee, C. Kim, C. Song, and W. Yu, "ANN control for damping low-frequency oscillation using deep learning," in *Proc. IEEE Aust. Univ. Power Eng. Conf.*, Auckland, New Zealand, 2018, pp. 1–4.
[31] J. Duan, H. Xu, and W. Liu, "Q-learning-based damping control of wide-area power systems under cyber uncertainties," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6408–6418, Nov. 2018.
[32] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015. [Online]. Available: arXiv:1509.02971.
[33] D. Ha and J. Schmidhuber, "World models," 2018. [Online]. Available: arXiv:1803.10122.
[34] C. J. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[35] L. Simon, K. S. Swarup, and J. Ravishankar, "Wide area oscillation damping controller for DFIG using WAMS with delay compensation," *IET Renew. Power Gener.*, vol. 13, no. 1, pp. 128–137, Jan. 2019.
[36] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[37] P. Kundur, N. J. Balu, and M. G. Lauby, *Power System Stability and Control*. New York, NY, USA: McGraw-Hill, 1994.
[38] M. Jin and J. Lavaei, "Stability-certified reinforcement learning: A control-theoretic perspective," 2018. [Online]. Available: arXiv:1810.11505.
[39] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Trans. Autom. Control*, vol. 42, no. 6, pp. 819–830, Jun. 1997.
[40] M. Klein, G. Rogers, S. Moorty, and P. Kundur, "Analytical investigation of factors influencing power system stabilizers performance," *IEEE Trans. Energy Convers.*, vol. 7, no. 3, pp. 382–390, Sep. 1992.
[41] T. Athay, R. Podmore, and S. Virmani, "A practical method for the direct analysis of transient stability," *IEEE Trans. Power App. Syst.*, vol. PAS-98, no. 2, pp. 573–584, Mar. 1979.

Yousuf Hashmy (Student Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, and the M.Sc. degree in electrical engineering from Arizona State University (ASU).

He was an Application Engineer in microgrid development technical cell with AESL-Caterpillar Inc., from November 2016 to December 2017. He served as a Research Assistant with the Ira A. Fulton Schools of Engineering, ASU. He is also trained in Energy Innovation and Emerging Technologies with Stanford University. He is currently a Research Associate with ITU, Lahore. He also served as a full-time Researcher with GEIRI North America. His research interests are in the areas of machine learning and big data applications in power systems, real-time control through reinforcement learning, power system protection, WAMS-based control systems, microgrid, global public and energy policy, and sustainable infrastructure for underdeveloped economies. He is a recipient of a Gold Medal from the Government of Pakistan for outstanding performance in academics.

Zhe Yu (Member, IEEE) received the B.E. degree from the Department of Electrical Engineering, Tsinghua University, Beijing, China, in 2009, the M.S. degree from the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, in 2010, and the Ph.D. degree from the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA, in 2016. He joined Global Energy Interconnection Research Institute North America (GEIRI North America) in 2017. His current research interests include power system and smart grid, machine learning, data mining, and optimization.

Di Shi (Senior Member, IEEE) received the B.S. degree in electrical engineering from Xian Jiaotong University, Xian, China, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2009 and 2012, respectively. He currently leads the AI and System Analytics Group, GEIRI North America, San Jose, CA, USA. His research interests include WAMS, energy storage systems, and renewable integration. He is an Editor of the IEEE TRANSACTIONS ON SMART GRID and the IEEE POWER ENGINEERING LETTERS.

Yang Weng (Member, IEEE) received the B.E. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, the M.Sc. degree in statistics from the University of Illinois at Chicago, Chicago, IL, USA, and the M.Sc. degree in machine learning of computer science and the M.E. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA.

He joined Stanford University, Stanford, CA, USA, as the TomKat Fellow for sustainable energy. He is currently an Assistant Professor of electrical, computer and energy engineering with Arizona State University, Tempe, AZ, USA. His research interest is in the interdisciplinary area of power systems, machine learning, and renewable integration. He received the CMU Dean's Graduate Fellowship in 2010, the Best Paper Award at the International Conference on Smart Grid Communication (SGC) in 2012, the First Ranking Paper of SGC in 2013, the Best Papers at the Power and Energy Society General Meeting in 2014, the ABB Fellowship in 2014, the Golden Best Paper Award at the International Conference on Probabilistic Methods Applied to Power Systems in 2016, and the Best Paper Award at IEEE Conference on Energy Internet and Energy system Integration in 2017, IEEE North American Power Symposium in 2019, and the IEEE Sustainable Power and Energy Conference in 2019.