Optimising Damping Control in Renewable Energy Systems through Reinforcement Learning within Wide-Area Measurement Frameworks

Truong Ngoc-Hung
Department of I.T., FPT University - Quy Nhon A.I Campus,
Nhon Binh ward, Quy Nhon city, Binh Dinh province, Vietnam
hungtn19@fe.edu.vn

***Abstract*—This paper introduces a reinforcement learning-based controller, utilising the deep deterministic policy gradient (DDPG) method, to mitigate low-frequency disturbances in electrical grids with renewable energy sources. It features a novel reward function inversely related to the control error and employs a state vector comprising absolute and integral errors to enhance error reduction. The controller, tested on a dual-region system with solar power, utilises phasor measurement unit (PMU) data for global inputs. Its performance is validated through time-domain simulations, pole-zero mapping, modal analysis, frequency response, and participation factor mapping, using a custom MATLAB and Simulink toolkit. The design accounts for communication delays and adapts to variable conditions, which proves to be effective in reducing oscillations and improving system stability.**

***Index Terms*—Reinforcement learning; Wide-area measurement systems; Deep deterministic policy gradient (DDPG); Solar plant; Inter-area oscillations.**

I. INTRODUCTION

The exponential rise in interconnected energy networks, along with the surge in renewable energy utilisation and the amalgamation of sources with low inertia and adaptable loads, has thrust inter-area fluctuations to the forefront as a pivotal issue for the stability of power grids. These fluctuations detrimentally influence the peak capacity of transmission lines and engage various elements of the energy network, leading to potential destabilisation [1], [2]. To tackle this issue, deploying effective control mechanisms is vital for quelling inter-area fluctuations and bolstering the resilience of energy infrastructures. The wide-area measurement system (WAMS) is instrumental in furnishing controllers with comprehensive data sourced from phasor measurement units (PMUs), characterised by their time-synchronised readings and high-frequency sampling.

In the realm of power system controls, two distinct methodologies are prevalent: model-centric and model-agnostic strategies. Model-centric approaches encapsulate the controller's functionality through algebraic formulations based on the system's model. The design of controllers within this paradigm for complex networks replete with model uncertainties demands an exhaustive physical representation and structure, a formidable task for vast interconnected systems characterised by extensive state and action dimensions [1]. In contrast, model-agnostic controllers are predicated on intelligent systems that hone their parameters by learning from the correlation between input and output data, obviating the need for a comprehension of an internal model or its mathematical representations, and instead leveraging real-time data and insights. This attribute renders model-agnostic controllers particularly apt for managing the intricacies of large-scale multifaceted power networks. The advent of sophisticated sensing technologies in smart grids has led to an influx of data, paving the way for the development of intelligent agents tailored to mitigate system deficiencies [3]. These controllers are designed to accommodate a spectrum of scenarios and operational dynamics, drawing on empirical insights to inform decision-making processes. Within the sphere of model-agnostic controllers, those founded on machine learning principles, especially reinforcement learning (RL), are regarded cutting edge. Machine learning encompasses three primary domains: supervised and unsupervised learning for static data categorisation and labelling, respectively, while RL is tailored for dynamic contexts, aiming to devise actions predicated on state observations and rewards accrued through ongoing interactions with the environment [4], [5].

Recent scholarly endeavours have extensively explored the application of RL within power systems, elucidating the development of intelligent controllers that specifically target low-frequency disturbances such as inter-area fluctuations [1], [6]–[12]. In one notable study, a robust controller based on WAMS is introduced [1], leveraging policy gradient techniques to modulate the field voltages of several synchronous generators. The reward function in this study integrates a variety of measurements, both remote and local, encompassing speed deviations and the persistence of relative speed variations, alongside voltage phase angle disparities at distant buses. The controller design is intricate, necessitating coordinated efforts across multiple controllers, each processing a collection of local and remote signals to mitigate a single problem inter-area fluctuations. However, this research predominantly concentrates on synchronous generators, with scant attention to the integration of solar facilities that lack a proven control scheme for damping.

Manuscript received 18 February, 2024; accepted 8 May, 2024.

Another study presents a multiband power system stabiliser that employs deep RL [6], with the controller's parameters fine-tuned through the proximal policy optimisation technique. This study presupposes a system exclusively composed of traditional generators, with controller inputs derived from local metrics such as rotor angles, active power, and voltage levels. Although numerous recent investigations have applied RL to tackle load frequency control issues, and there are extensive reviews on power system management via RL, there is a notable paucity of focus on inter-area fluctuations within networks incorporating renewable energy sources [13]–[16]. Inclusion of low-inertia resources in these networks decreases damping capabilities, thereby exacerbating instability, complexity, and unpredictability.

This study introduces a control strategy based on RL, utilising the deep deterministic policy gradient DDPG methodology, to counteract inter-area oscillations. The control mechanism is positioned on the side devoid of inertia, aligned with a solar energy installation, and it procures its remote-signal input from the WAMS. The demonstration system is structured around a dual-area setup, with all power system stabilisers (PSSs) excluded to underscore the efficacy of the newly proposed controller in damping oscillations. Extensive analysis and simulation of the system are executed, taking into account time delays, using specialised software designed for this investigation. The paper addresses low-frequency oscillations by merging leading-edge technologies, incorporating WAMS, machine learning, and renewable energy through a photovoltaic facility, thus eliminating the need for additional damping mechanisms, such as stabilisers. Remarkably, the proposed controller employs dual error signals to swiftly rectify the implemented action [3], [17]–[21]. Recognising the potential for communication delays in global measurements, the controller design encompasses a wide spectrum of practical latency intervals. All software, algorithms, and graphical representations delineated in this study are the exclusive creations of the author.

The ensuing segments of this paper are organised as follows. Section II outlines the control scheme enacted by the RL agent. The problem statement is delineated in Section III, followed by an introduction to the RL framework in Section IV. The test system is elaborated in Section V, Section VI presents the findings and begins a discussion, culminating in Section VII, which provides the concluding remarks.

II. REINFORCEMENT LEARNING CONTROL

In traditional feedback control systems, engineers usually use adaptive or optimal control strategies. The key difference between these approaches lies in the way controller parameters are adjusted: either in real-time (online) or prior to deployment (offline). Adaptive control dynamically fine-tunes parameters based on ongoing measurements without seeking to optimise them [22]. On the contrary, optimal control involves pre-setting parameters through mathematical optimisation, which requires a detailed model of the system's dynamics. Reinforcement learning (RL) introduces a machine learning paradigm to understand a system response to inputs and outputs, facilitating the creation of controllers that blend adaptability with optimisation. Unlike conventional techniques, RL bypasses the need for direct system modelling, instead adjusting controller parameters via a learning process based on data. RL consists of two main elements: the agent, which represents the controller under development, and the environment, which includes the entire system apart from the controller. In this setup, the agent and environment are linked by a feedback loop involving a control signal from the agent to the environment and two signals from the environment to the agent. These signals deliver the system output as feedback and a reward function, which evaluates the effectiveness of the agent's actions. Within the agent, two neural networks, known as the actor and the critic, require optimisation of their parameters to maximise the information gleaned about the behaviour of the system. Figure 1 of the original text visually depicts these concepts, with Fig. 2 illustrating the theoretical framework and the other showing a practical implementation in Simulink.

The agent applies a specific algorithm to define the final policy of the controller. A common choice in RL is the deep deterministic policy gradient (DDPG), an actor-critic method suitable for a range of action spaces, both continuous and discrete. The DDPG involves two models: the actor, which maps states to actions, and the critic, which evaluates the proposed actions based on a combination of states and actions through a Q-value. This Q-value helps to determine the desirability of the actions suggested by the actor. The actor is often termed the policy maker, whereas the critic is known as the policy evaluator.

This research aims to address the challenge of reducing inter-area oscillations in a solar plant control loop by incorporating a stabilising signal, using data from wide-area measurement systems (WAMS) to detect and counteract oscillations. The goal is to develop an agent capable of modifying the control loop to effectively mitigate these oscillations within a specified environment that includes the state of the system and communication delays.

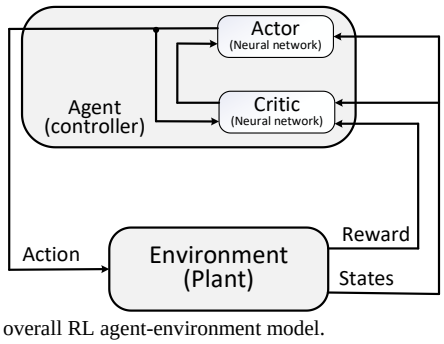


Fig. 1. The overall RL agent-environment model.

Figure 1 illustrates the interaction between an Agent (controller) and its Environment (Plant) in a reinforcement learning framework, specifically focussing on an Actor-Critic architecture. This setup represents an Actor-Critic model, which is a popular approach in RL. The actor and the critic work together to improve the agent's policy, where the actor focusses on exploring actions and the critic ensures that these actions lead to optimal long-term rewards.

1. Environment (Plant): - The environment represents the system or context within which the agent operates. This could be anything from a simulated environment to a physical system in real-world applications. - States: The environment provides the current state information to the agent, representing the situation or configuration of the

environment. - Reward: The environment gives a reward to the agent based on the action taken, which serves as feedback on the performance of the action. - Action: The environment receives the action decided by the agent and changes its state accordingly.

2. Agent (Controller): - The agent is the entity responsible for making decisions and learning from interactions with the environment to optimise its performance. - The agent here is composed of two main components: + Actor (Neural network): The actor is responsible for selecting actions based on the current state. It uses a neural network to determine the best action to take in each state. + Critic (Neural network): The critic evaluates the actions taken by the actor by estimating the value function, which assesses the expected cumulative reward from each state. It provides feedback to the actor to improve future action selections.

3. Interactions: - *Action*: The actor network within the agent selects an action based on the current state of the environment. - *States and Reward*: The environment transitions to a new state and provides a reward based on the action taken by the agent. - *Critic Feedback*: The critic evaluates the action by comparing the predicted value of the current state-action pair to the actual received reward and updates its value function accordingly. This feedback helps the actor refine its policy to choose better actions in the future.

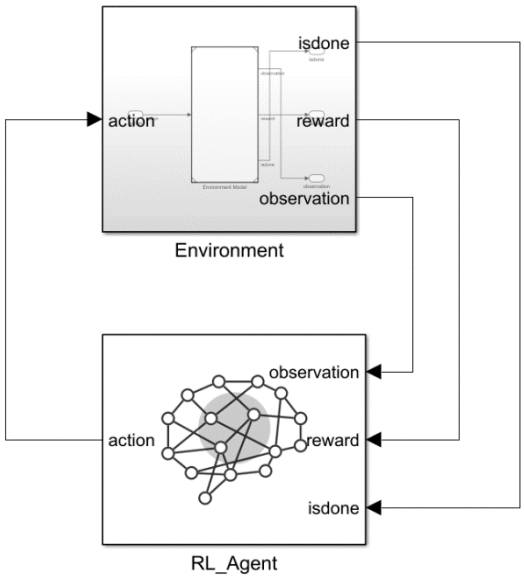


Fig. 2. Simulink diagram of the agent-environment model.

Figure 2 shows the interaction between an RL Agent and its Environment, showcasing the fundamental loop of a reinforcement learning system.

1. Environment: - This represents the external system or context in which the agent operates. The environment can be anything from a simulated game to a real-world scenario such as robotic control. -Action: The agent takes an action based on its policy or decision-making process. - Observation: After taking the action, the agent receives an observation from the environment. This observation includes information about the current state of the environment. - Reward: The agent also receives a reward, which is a numerical value that indicates immediate feedback on the effectiveness of the action. - IsDone: This signal indicates whether the episode (a complete sequence of actions that lead to the final state) has ended.

2. RL_Agent: - The agent represents the learning entity that interacts with the environment to maximise cumulative rewards over time. - Observation: The agent receives observations from the environment to understand the current state. - Reward: The agent receives rewards to evaluate the immediate outcome of its actions. - IsDone: This signal informs the agent if the current episode is over, helping it determine when to reset and start a new episode. - Action: Based on the observation and reward received, the agent decides on the next action to take, aiming to maximise future rewards. The loop continues with the agent constantly interacting with the environment through this cycle, learning, and improving its policy based on the feedback received.

III. REINFORCEMENT LEARNING AGENT

In this research, we introduce a control mechanism based on RL to effectively counteract inter-area disturbances, leveraging a distant signal from a comprehensive measurement network. The controller's decisions are influenced by its inputs, specifically the choice of certain data points as state observations to improve its grasp of such disturbances when they manifest. The concept of inter-area oscillation, characterised by the synchronous movement of machine groups across different regions, is addressed by aggregating measurements from each region to create a differential centre-of-inertia. Frequency data from phasor measurement units (PMUs) are utilised due to their critical role and responsiveness to system fluctuations. The system configuration, including the RL controller with dual inputs for observation and reward, along with an episode termination indicator called "*isdone*", is depicted in a Simulink diagram in Fig. 3.

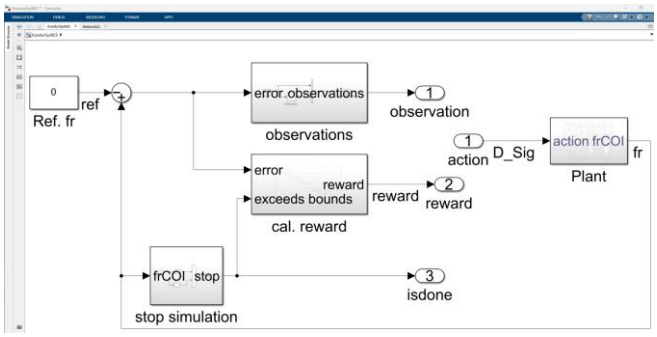


Fig. 3. Control scheme of agent-environment using Simulink.

Figure 3 appears to illustrate a Simulink model for a reinforcement learning setup, focussing on the interaction between the components of the system. The model illustrates a typical reinforcement learning loop where the agent interacts with the environment (plant), receives observations and rewards based on its actions, and decides on subsequent actions to achieve the desired reference input (*Ref. fr*). The simulation continues until the stop condition is met, as indicated by the *isdone* signal.

1. Ref. fr: This is likely the reference input or the setpoint that the system aims to achieve. It could represent the desired state or output of the system.

2. Observations: - This block processes the error (difference between the reference input and the current state) and generates observations. - The *error.observations* signal represents the processed observations that are used to understand the current state of the system. - Output observation ①: This signal is sent to the agent to inform it about the current state of the environment.
3. Reward Calculation (cal. reward): - This block calculates the reward based on the error and whether it exceeds certain bounds. - The error exceeds bounds signal likely indicates if the system performance is outside acceptable limits. - Output reward ②: This reward signal is provided to the agent as feedback on its action.
4. Stop Simulation (frCOI.stop): - This block determines if the simulation should be stopped. - The *isdone* ③ signal indicates whether the current episode or simulation run has ended.
5. Agent: - The agent, likely implemented in a separate block or as part of a reinforcement learning algorithm, receives the observation ① and reward ② signals. - It processes these inputs to decide the next action. - Output action *D_Sig*: This is the action signal generated by the agent based on the observation and reward received, which is then sent to the environment.
6. Plant: - The plant represents the environment or the system being controlled. - It receives the action *frCOI* signal (denoted as action *D_Sig* in the agent block), which influences its state. - The output *fr*: This is the current state or output of the plant, which is fed back to the observations and reward calculation blocks to close the loop.

A. Observation Signal

In typical control systems, a feedback loop is established by comparing the system's output to the input reference. Similarly, in the construction of an RL-based controller, a state vector derived from the system output, here the differential frequency deviations, serves as the observation signal in Fig. 3. This signal acts as an error input, with its integral added to introduce a temporal memory component, which helps to reduce cumulative errors, as illustrated in subsequent Fig. 4.

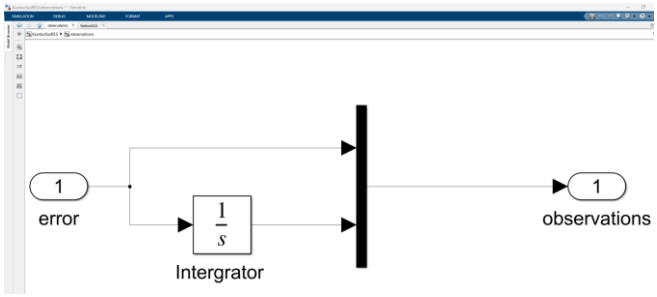


Fig. 4. Observation signal.

B. Reward Function

Training an RL agent is a complex task, often requiring a long time to derive meaningful insights from the data and develop an effective strategy. Enhancing the reward function is a tactic to accelerate the agent's learning process. A meticulously designed reward function is vital to accelerate the agent's learning curve. In the depicted control scheme, the agent receives a reward inversely related to the error magnitude, incentivising lower errors with higher rewards to minimise deviations and improve tracking accuracy. A unique discrete reward mechanism is employed, adjusting rewards based on the error magnitude within a predefined range in Fig. 5.

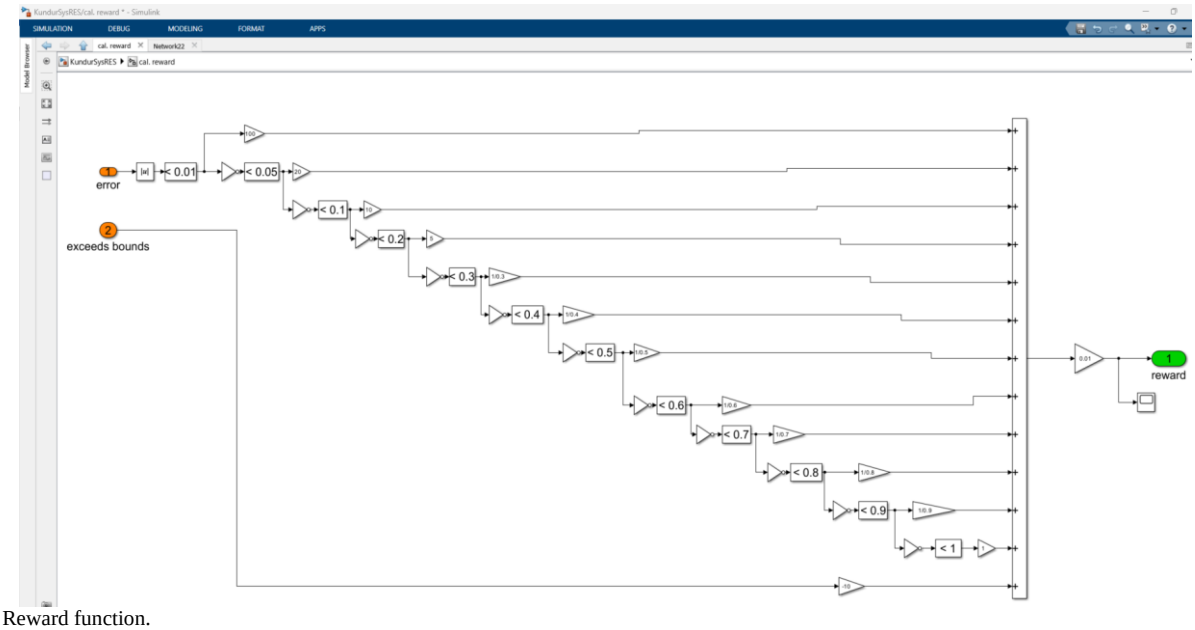
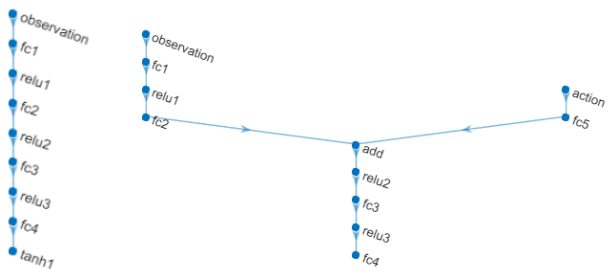


Fig. 5. Reward function.

C. Actor-Critic Networks

The framework involves constructing two deep neural networks for the actor and critic components to forecast the cumulative reward. The architecture of these networks, including the number of neurons and layers, is determined through trial and error, with a focus on fully connected layers. The actor network processes the state vectors, including the error and its integral, through multiple layers, while the critic network evaluates the actor's decisions. To optimise learning, normalisation techniques and activation functions are applied across the networks, with initial weights set close to zero in Fig. 6.



(a) (b)

Fig. 6. (a) Actor network; (b) Critic network.

D. *Training Parameters*

The training regimen for the agent encompasses specific parameters, such as action signal limitations corresponding to generator capabilities, defined sampling intervals, and simulation durations. A reset function ensures consistent initial conditions for each simulation run, with the training duration and episode count tailored to achieve desired damping effects according to the defined reward structure. Additional training considerations include the discount factor, which indicates the value of future rewards, and noise parameters to prevent overfitting and improve generalisation of the learning process.

IV. ENVIRONMENT

In this investigation, we have engineered an agent specifically to counteract a prevalent challenge within the system, known as inter-area oscillations. Our focus is on the power system, excluding the agent from the environment under consideration. We employed Simulink to both design the agent and simulate the power system, following methodologies outlined in previous work [2]. The intricate sixth-order generator model, complete with excitation and turbine mechanisms, is adopted as shown in one of Fig. 7.

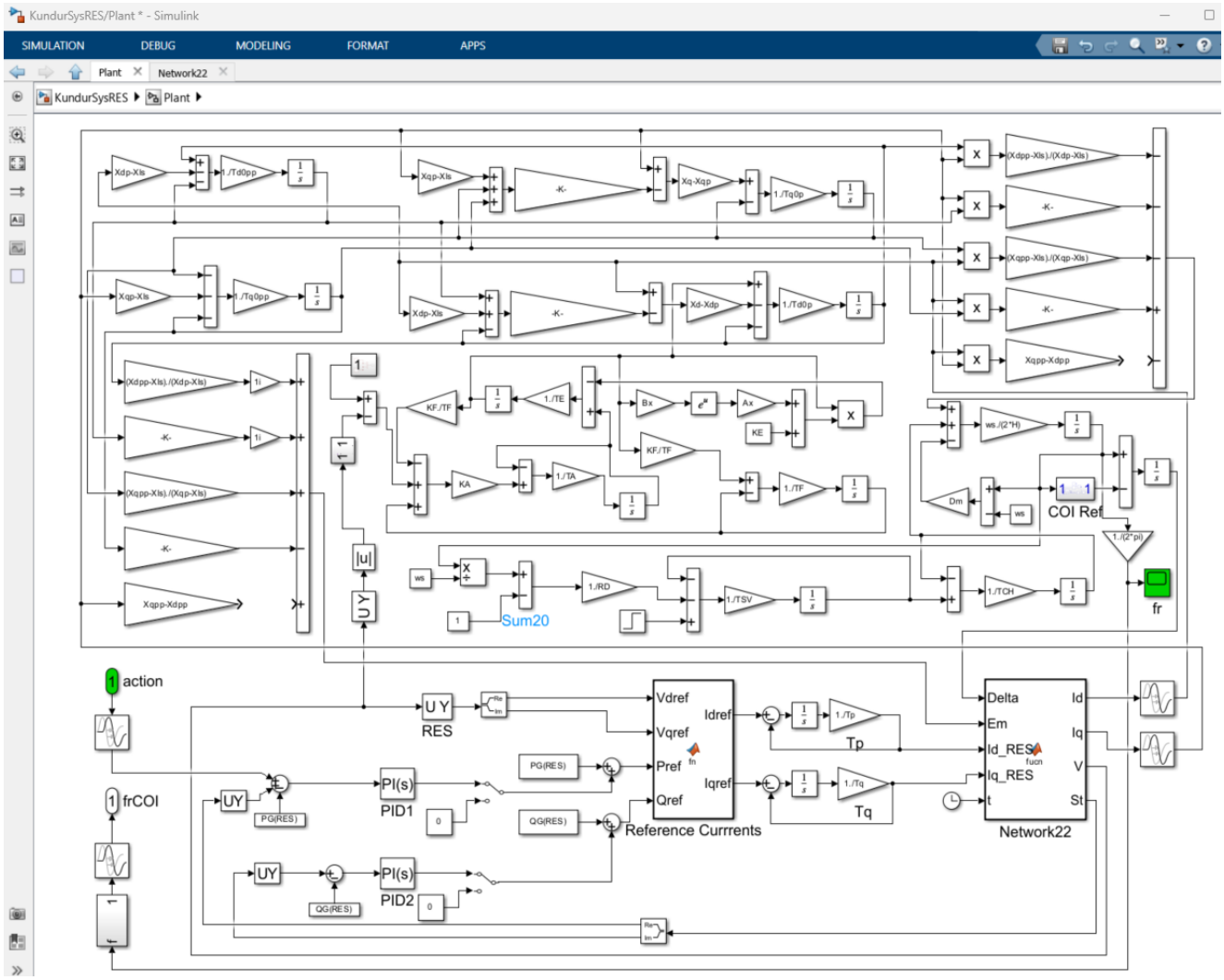


Fig. 7. The environment (plant) diagram implemented in Simulink.

In Peter Sauer's "Power System Dynamics and Stability", several key concepts are elaborated that are critical to understanding the behaviour and control of power systems. On page 176, Sauer delves into the intricacies of transient stability analysis, emphasising the importance of accurate modelling in predicting system response to disturbances [2]. The discussion on page 109 highlights the fundamental principles of rotor angle stability, illustrating the dynamic interactions between synchronous machines and the electrical network. Furthermore, page 111 explores voltage stability, outlining the conditions under which a power system can maintain acceptable voltage levels following significant changes in load or generation. Together, these sections provide a comprehensive overview of the dynamic challenges faced in maintaining stable and reliable power system operation.

Equations (1) to (13) illustrate a detailed Simulink model showing the control and management system for a renewable energy source integrated with a wide-area measurement system in Fig. 7.

$$\bar{Y}_{aug} = \begin{array}{c} \\ n+1 \\ \cdots \\ n+m \\ 1 \\ \cdots \\ m \\ m+1 \\ \cdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} n+1\ldots n+m & & 1\ldots m & & m+1\ldots n \end{array} \\ \begin{pmatrix} & & & \\ \bar{y} & & -\bar{y} & 0 \\ & & & \\ & & & \\ -\bar{y} & & & \\ & & & \\ & & \bar{y}_{N1} & \\ & & & \\ 0 & & & \\ & & & \end{pmatrix} \end{array}, \quad (1)$$

where $\bar{y} = Diag\left(\frac{1}{jX'_{di}}\right), \quad i = 1,...,m$ and $\bar{y}_{N1} = \bar{Y}_N + \begin{bmatrix} \bar{y} & 0 \\ 0 & 0 \end{bmatrix}$.

This type of matrix is particularly relevant in dynamic simulations and analyses, where understanding the interdependencies within a network is crucial. For example, in power systems, this could help in stability analysis, fault analysis, and design of control strategies to ensure reliable and efficient operation. The specific arrangement and the presence of zeros also suggest optimisations in computational analyses by reducing the complexity of interactions that need to be considered.

$$\left| \begin{array}{l} \left(X''_{qi} - X''_{di}\right) I_{qi} + \frac{\left(X''_{qi} - X_{\ell si}\right)}{\left(X'_{qi} - X_{\ell si}\right)} E'_{di} - \\ -\frac{\left(X'_{qi} - X''_{qi}\right)}{\left(X'_{qi} - X_{\ell si}\right)} \psi_{2qi} + j \frac{\left(X''_{di} - X_{\ell si}\right)}{\left(X'_{di} - X_{\ell si}\right)} E'_{qi} + \\ + j \frac{\left(X'_{di} - X''_{di}\right)}{\left(X'_{di} - X_{\ell si}\right)} \psi_{1di} \end{array} \right| e^{j\left(\delta_i - \pi/2\right)}. \quad (2)$$

This equation is part of the foundational analysis to ensure the reliability and efficiency of power transmission and distribution networks, especially in scenarios that challenge the stability of the system.

This representation gives the following dynamic model for the *m* machine, *n* bus power system after the stator/network and load electrical transients have been eliminated, and using the load model.

$$T'_{doi} \frac{dE'_{qi}}{dt} = -E'_{qi} - (X_{di} - X'_{di}) \times \times \left[I_{di} - \frac{\left(X'_{di} - X''_{di}\right)}{\left(X'_{di} - X_{\ell si}\right)^2} \begin{pmatrix} \psi_{1di} + \\ + \left(X'_{di} - X_{\ell si}\right) I_{di} - E'_{qi} \end{pmatrix} \right] + E_{fdi}, \quad \{i = 1,...,m\}, \quad (3)$$

where $T'_{doi} \frac{dE'_{qi}}{dt}$ represents the time derivative of E'_{qi}, scaled by a time constant T'_{doi}. E'_{qi} is usually the internal voltage behind the transient reactance of the i[th] machine; X_{di}: d-axis synchronous reactance; X'_{di}: d-axis transient reactance; X''_{di}: Some specific reactance value, possibly a steady-state value, another specific reactance, possibly a nominal value; I_d: The direct axis current component; ψ_{1di}: This could be a flux linkage or another electrical parameter related to the d-axis; E'_{qi}: Another voltage term likely related to the i[th] generator or machine. This differential equation captures the dynamics of the internal voltage E'_{qi} in a synchronous machine, taking into account various reactances, currents, and other electrical parameters that influence the machine's behaviour over time; ℓ : Total number of loops in the graph (or circuit); *m*: Number of synchronous machines (if there is an infinite bus, it is number 1); *n*: Number of three-phase system buses (excluding the datum or references bus); *b*: Total number of machines plus transformers plus lines plus loads (total branches).

$$T''_{doi} \frac{d\psi_{1di}}{dt} = -\psi_{1di} + E'_{qi} - (X_{di} - X'_{\ell si}) I_{di}, \quad \{i = 1,...,m\}, \quad (4)$$

where $T''_{doi} \frac{d\psi_{1di}}{dt}$ represents the time derivative of ψ_{1di} scaled by a time constant T''_{doi}. ψ_{1di} : The flux linkage on the d-axis for the i[th] machine; E'_{qi} : The internal voltage behind the transient reactance for the i[th] machine, as discussed in the previous equation; X_{di} and X'_{qi}: Axis synchronous reactance and q-axis transient reactance, respectively; I_{di}: The current component of the i-axis, which might be another axis used in this specific model of the synchronous machine. This equation, combined with the previous one, helps to describe the behaviour of electrical parameters within a synchronous machine over time, considering various reactances, currents, and voltages.

$$T'_{qoi} \frac{dE'_{qi}}{dt} = -E'_{di} + (X_{qi} - X'_{qi}) \times \times \left[I_{qi} - \frac{\left(X'_{qi} - X''_{qi}\right)}{\left(X'_{qi} - X_{\ell si}\right)^2} \begin{pmatrix} \psi_{2qi} + \\ + \left(X'_{qi} - X_{\ell si}\right) I_{qi} + E'_{di} \end{pmatrix} \right], \quad \{i = 1,...,m\}, \quad (5)$$

where $T'_{qoi} \frac{dE'_{qi}}{dt}$ represents the time derivative of E'_{qi}, scaled by a time constant T'_{qoi}. E'_{qi} is usually the internal voltage behind the transient reactance of the i[th] machine on the q-axis; E'_{di}: This term represents the internal voltage behind the transient reactance for the i[th] machine on the d-axis; X_{qi}: q-axis synchronous reactance; X'_{qi}: q-axis transient reactance; I_{qi}: q-axis current component; ψ_{2qi}: This could be a flux linkage or another electrical parameter related to the q-axis; X''_{qi} and X_{lsi}: These terms represent nominal and steady-state reactance values for the q-axis, respectively; E'_{di}: Another voltage term related to the i[th] generator or machine on the d-axis.

$$T''_{qoi}\frac{d\psi_{2qi}}{dt}=-\psi_{2qi}-E'_{di}-(X'_{qi}-X_{\ell si})I_{qi},\quad \{i=1,...,m\},\tag{6}$$

$$\frac{d\delta_i}{dt}=\omega_i-\omega_s,\qquad \{i=1,...,m\},\tag{7}$$

$$\frac{2H_i}{\omega_s}\frac{d\omega_i}{dt}=T_{Mi}-\frac{\left(X''_{di}-X_{\ell si}\right)}{\left(X'_{di}-X_{\ell si}\right)}\mathrm{E}'_{qi}\mathrm{I}_{qi}-\frac{\left(X'_{di}-X''_{di}\right)}{\left(X'_{di}-X_{\ell si}\right)}\psi_{1di}\mathrm{I}_{qi}-$$
$$-\frac{\left(X''_{qi}-X_{\ell si}\right)}{\left(X'_{di}-X_{\ell si}\right)}\mathrm{E}'_{di}\mathrm{I}_{di}+\frac{\left(X'_{qi}-X''_{qi}\right)}{\left(X'_{qi}-X_{\ell si}\right)}\psi_{2qi}\mathrm{I}_{di}-$$
$$-\left(X''_{qi}-X''_{di}\right)\mathrm{I}_{di}\mathrm{I}_{qi}-\mathrm{T}_{FWi},\qquad \{i=1,...,m\},\tag{8}$$

where H_i: Inertia constant of the ith generator; ω_s: Synchronous angular velocity; ω_i: Angular velocity of the ith generator; T_{Mi}: Mechanical torque input to the ith generator; X''_{di}, X''_{qi}: Transient reactance of the ith generator along the d-axis and q-axis, respectively; X_{lsi}: Leakage reactance of the ith generator; X_{ti}: Total reactance of the ith generator; E'_{qi}, E'_{qi}: Transient voltages of the ith generator along the q-axis and d-axis, respectively; I_{di}, I_{qi}: Current components of the ith generator along the d-axis and q-axis, respectively; V_{di}, V_{2qi}: Voltage components along the d-axis and q-axis, respectively; T_{FI}: Friction and windage losses or torque. This equation requires a good grasp of power system dynamics, particularly the interaction between mechanical and electrical quantities in synchronous generators.

$$T_{Ei}\frac{dE_{fdi}}{dt}=-\left(K_{Ei}+S_{Ei}\left(E_{fdi}\right)\right)E_{fdi}+V_{Ri},\quad \{i=1,...,m\},\tag{9}$$

where T_{Ei} : Time constant of the excitation system for the ith generator; E_{fdi}: Field voltage of the ith generator; K_{Ei}: Gain of the excitation system for the ith generator; $S_{Ei}(E_{fdi})$: Saturation function of the excitation system, which is a function of the field voltage E_{fdi}; V_{Ri}: Reference voltage input to the excitation system for the ith generator; i: Index representing the generator (from 1 to m). This equation is part of a set of differential equations that describe the behaviour of generators in a power system, capturing the interactions between mechanical, electrical, and control components.

$$T_{Fi}\frac{dR_{fi}}{dt}=-R_{fi}+\frac{K_{Fi}}{T_{Fi}}E_{fdi},\qquad \{i=1,...,m\},\tag{10}$$

where T_{Fi}: Time constant of the field winding for the ith generator; R_{fi}: Rotor field current of the ith generator; K_{Fi}: Gain factor associated with the field winding of the ith generator; E_{fdi}: Field voltage of the ith generator; i: Index representing the generator (from 1 to m). This equation, along with others in the system, provides a comprehensive model of the dynamic behaviour of the generator, allowing engineers to predict and mitigate potential stability issues in the power grid.

$$T_{Ai}\frac{dV_{Ri}}{dt}=-V_{Ri}+K_{Ai}R_{fi}-\frac{K_{Ai}K_{Fi}}{T_{Fi}}E_{fdi}+$$
$$+\mathrm{K}_{Ai}\left(V_{\mathrm{refi}}-V_i\right),\qquad \{i=1,...,m\},\tag{11}$$

$$T_{CHi}\frac{dT_{Mi}}{dt}=-T_{Mi}+P_{SVi},\qquad \{i=1,...,m\},\tag{12}$$

where $T_{CHi}\frac{dT_{Mi}}{dt}$ is the rate of change of the parameter T_{Mi} with respect to time modulated or scaled by another parameter, T_{CHi}. This could suggest that T_{CHi} influences how quickly T_{Mi} changes, possibly representing a capacity or inertial effect in thermal or process dynamics; The equation could be found in thermal management systems, where T_{Mi} might represent temperatures in different parts of the system, with $+P_{STi}$ possibly being power inputs or other energetic contributions, and T_{CHi} representing the heat capacity or thermal inertia of each component.

$$T_{SVi}\frac{dP_{SVi}}{dt}=-P_{SVi}+P_{Ci}-\frac{1}{R_{Di}}\left(\frac{\omega_i}{\omega_s}-1\right),\quad \{i=1,...,m\},\tag{13}$$

where $T_{SVi}\frac{dP_{SVi}}{dt}$ term suggests a product of a temperature-like parameter T_{STi} and the rate of change of pressure P_{STi} with respect to time. This could be indicative of a thermal energy change or flow dynamics within a system.

$-P_{STi}$ + P_{Ci}: This part involves the subtraction of two pressures, P_{STi} and P_{Ci}. This difference might represent a driving force or a potential drop at two points in a system.

$\frac{1}{R_{Di}}\left(\frac{\omega_i}{\omega_s}-1\right)$: R_{Di} could represent a resistance or a similar parameter that resists flow or change; $\frac{\omega_i}{\omega_s}$ suggests a ratio of two quantities, perhaps areas, coefficients, or other properties that vary with time or position, indicating how the system's behaviour deviates from some initial or reference state. The equation suggests that it models the pressure dynamics in a multicomponent system where the behaviour of each component is affected by its own characteristics (such as temperature and resistance) and its difference in pressure with another point in the system.

Figure 7 illustrates a detailed Simulink model showing the control and management system for a renewable energy source integrated with a wide-area measurement system. The model likely focusses on optimising damping control using reinforcement learning techniques, as inferred from the context.

1. Top Section (Transformers and Mathematical Operations): *1.1* - The top part consists of a series of transformers, mathematical blocks (adders, multipliers), and signal flow paths; *1.2* - Blocks such as "$\mathrm{X_{dp}}$ + $\mathrm{X_d}$" and "$\mathrm{X_{dp}}$ - $\mathrm{X_d}$" are seen, indicating operations on variables related to reactance's; *1.3* - There are various integrator blocks and gains, showing dynamic calculations for system variables.

2. Middle Section (Control Logic and Summation): *2.1* - This section has numerous interconnected blocks that probably represent control logic and dynamic response calculation; *2.2* - The blocks include gain blocks, integrators, summation points, and possibly some

proportional-integral-derivative (PID) controllers or filters; *2.3* - The presence of summation points (e.g., "Sum20") indicates the aggregation of different input signals.

3. Bottom Section (Current and Voltage Reference Generation): *3.1* - The bottom section appears to focus on generating reference currents and voltages; *3.2* - Blocks labelled "Reference Currents", "I_{dref}", "I_{qref}", "V_{dref}", and "V_{qref}" are prominent, suggesting control references for direct and quadrature axis currents and voltages; *3.3* - The lower part shows interconnections between PI controllers, signal multipliers, and trigonometric function blocks.

4. Inputs and Outputs: *4.1* - The model has various input blocks (such as "action" and "frCOI") and output connections indicating where signals enter and leave the system; *4.2* - The outputs include variables such as "I_d", "I_q", "V", and "S_t", representing current, voltage, and status signals.

5. Interconnected Systems: *5.1* - The block"Network22" at the bottom indicates a subsystem or another network model integrated within this model; *5.2* - There are several feedback loops, which implies the presence of control loops to manage system stability and performance.

6. Specific Blocks and Notations: *6.1* - The "PI(s)" and "PID" blocks are visible, indicating the presence of PI and PID controllers; *6.2* - Blocks like "COI Ref" and "COI Ret" are seen, which might be related to centre of inertia reference signals; *6.3* - The notation "U_Y RES" and "QG(RES)" suggest the handling of variables of the renewable energy system.

The theoretical framework of the system, encapsulated by differential-algebraic equations [23], is thoroughly elaborated in another study, which also explains how to establish the initial conditions for various parameters and states. The modelling of the solar power facility leans on a dual-loop control scheme recommended by a renewable energy group, a model that has gained acceptance among scholars and is a staple in many advanced software tools for dynamic assessments of extensive power networks. Load modelling amalgamates constant power, voltage, and impedance aspects. It is crucial to standardise the phase angles throughout the system [24], which can be achieved by adjusting the speeds relative to the velocity of a machine or the collective inertia of the system in Fig. 8.

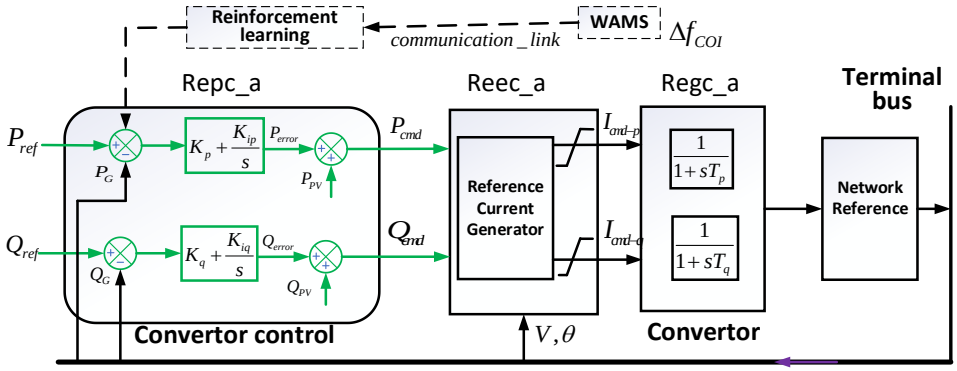


Fig. 8. Control loops.

In Fig. 8, a comprehensive control strategy is described for a power converter in a renewable energy system, integrating reinforcement learning and wide-area measurements to enhance stability and efficiency. The control strategy involves reference power adjustments, current generation, and dynamic regulation of the converter outputs to match network requirements.

1. Reinforcement Learning Integration: - The upper left corner of the diagram shows an input from a reinforcement learning model. This model provides control signals (denoted as \(\Delta P_{ref}\) and \(\Delta Q_{ref}\)) to the converter control system. - These control signals are modifications or adjustments to the reference power values \(P_{ref}\) and \(Q_{ref}\).

2. WAMS (Wide-Area Measurement Systems): - The diagram indicates the integration of WAMS, which provides wide-area measurements, \(\Delta f_{COI}\). These data help adjust the control signals for better stability and system performance.

3. Converter Control (Repc_a): - The first block labelled "Converter Control" (Repc_a) receives inputs \(P_{ref}\) and \(Q_{ref}\), which are the reference active and reactive power, respectively. - It includes two PI controllers to regulate the power outputs: - One PI controller for active power \(P_{ref}\). - One PI controller for reactive power \(Q_{ref}\). - The outputs of these PI controllers are denoted as \(v_d\) and \(v_q\), representing voltage components in the d-q reference frame.

4. Reference Current Generator (Recc_a): - The second block is the "Reference Current Generator" (Recc_a). - It takes the voltage components \(v_d\) and \(v_q\) and converts them into reference currents \(i_d^{*}\) and \(i_q^{*}\). - This block is crucial for generating the appropriate current references needed for the converter.

5. Converter (Regc_a): - The third block labelled "Converter" (Regc_a) receives the reference currents \(i_d^{*}\) and \(i_q^{*}\). - This block contains a controller that regulates the actual converter currents to match the reference currents. - The controller uses a transfer function \(\frac{1}{T_s}\) to model the dynamics of the converter system.

6. Network Reference: - The final block labelled "Network Reference" represents the connection to the terminal bus and the broader network. - It ensures that the converter output is synchronised with the network requirements and references.

7. Signal Flow and Connections: - Arrows indicate the flow of signals between blocks. - The entire system shows

a closed-loop control mechanism where the reinforcement learning model, converter control, and network references interact to maintain optimal performance.

We depict an intricate overview of the RL agent within its operational context, detailing the interconnections [25]–[27] and the nuanced architecture of the neural networks that embody the components of the actor and the critic, highlighted in red and green, respectively. The visual representation includes nodes and connections, with specific activation functions marked to clarify the network operational dynamics in Fig. 9.

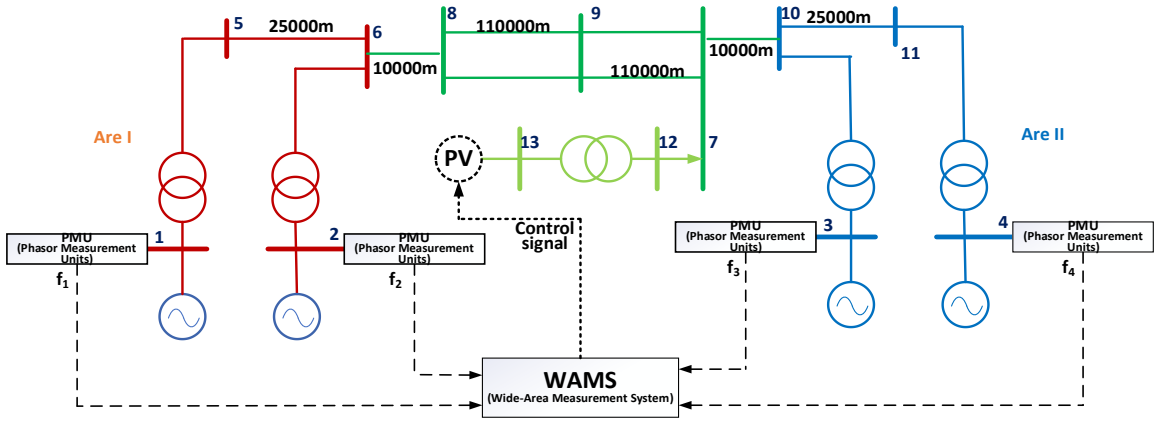


Fig. 9. A two-area test system.

V. TEST SYSTEM AND DEVELOPMENT OF MATLAB/SIMULINK

To develop and assess the designed agent, we employed a two-region test setup featuring a solar energy installation in one of the areas. The controller, situated in the solar facility, adds extra damping to the active power channel to efficiently dampen oscillations. The foundational data for this test arrangement were derived from the specific literature [2]. We deliberately omitted power system stabilisers (PSSs) from the setup to solely examine the effectiveness of the controller to dampen oscillations without other influences [27]. In the initial setup, an active power transfer of 400/900 per unit (pu) was anticipated across the interconnecting line. The generation of conventional generators in the first area was decreased by 200/900 pu, with the solar installation compensating for this reduction. Despite these changes, overall power flow was maintained. A schematic showing the integration of the solar facility into the network is illustrated in Fig. 10.

To delve into this issue, incorporate the solar power system, create the control mechanism, and conduct simulations in both the time and frequency domains, a comprehensive suite of tools and scripts was developed in MATLAB/Simulink. The schematic of the primary simulation model showcases the different system components like generators, exciters, turbines, and the solar facility, each marked in unique colours for clarity. Additional aspects of the system, such as the control unit, reward mechanism, state vector, and reference for tracking, are depicted in Fig. 7. These tools and scripts are designed for extended research and are openly provided for educational and research use.

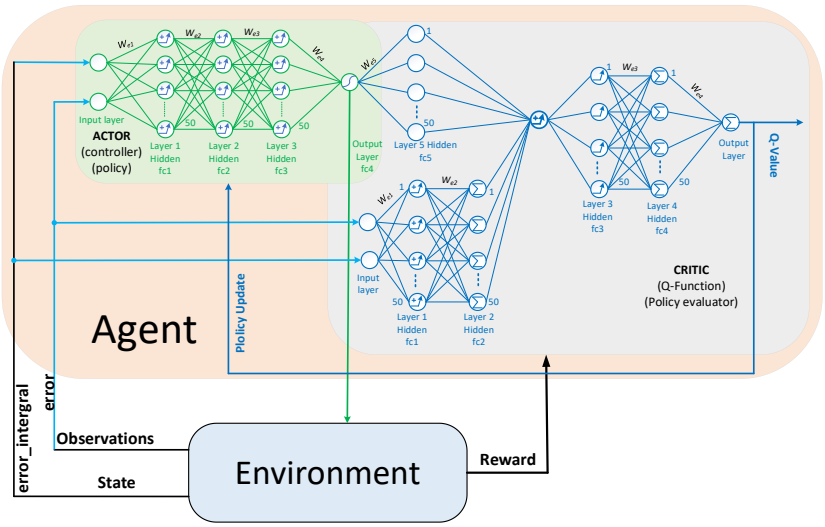


Fig. 10. Complete diagram of the system, including the environment and agent.

Figure 10 illustrates a reinforcement learning setup where an Agent interacts with its Environment, using an Actor-Critic architecture to learn and optimise its actions based on the feedback received. The diagram highlights the dynamic process of learning through continuous interaction and adjustment of the policy to maximise rewards.

1. Main Components: The diagram is divided into two main sections: the "Agent" and the "Environment".

2. Agent Section: *2.1* - Actor-Critic Model: The agent is shown to use an Actor-Critic architecture: *2.1.1* - Actor:

This part of the agent is responsible for selecting actions based on the current policy. Receive input signals (observations from the environment) and process them through a series of neural networks. The output is an action; *2.1.2* - Critic: This part evaluates the action taken by the actor. It estimates the value function, which helps in improving the policy by providing feedback (rewards); *2.2* - Network Layers: The agent structure includes multiple layers of the neural network, represented by interconnected nodes: *2.2.1* - Policy Network: Processes input to generate actions; *2.2.2* - Value Network: Evaluates actions to determine the expected reward.

3. Environment Section: *3.1* - The environment interacts with the agent by providing observations and rewards: *3.2* - Observations: The state of the environment, which is sent to the agent as input; *3.3* - Rewards: Feedback from the environment based on the actions taken by the agent is used to adjust the policy.

4. Interactions: *4.1* - State and Observations: The environment provides the state and observations to the agent; *4.2* - Action: The agent sends actions back to the environment based on its policy; *4.3* - Reward: The environment provides rewards to the agent, which are used for learning and improving the policy; *4.4* - These interactions form a loop where the agent continuously learns and adapts its policy to maximise the cumulative reward over time.

5. Detailed Elements: *5.1* - *ACER (Actor-Critic with Experience Replay):* Indicates that the agent uses an experience replay mechanism to improve learning efficiency; *5.2* - *Gated Recurrent Units (GRUs):* May be part of the neural network architecture to handle sequential data; *5.3* - *Q-Function:* Represents the expected reward of taking an action in a given state, used by the critic to evaluate the policy.

6. Connections and Flow: *6.1* - Arrows indicate the flow of information between different components; *6.2* - The blue and green lines show the paths for different signals (state, action, reward, observations).

VI. Results and Discussion

To understand the problem more deeply, a comparative study of the behaviour of the system with and without the incorporation of a control mechanism was carried out through time-domain simulations, modal examinations, participation factor analysis, and assessments of frequency response. The foundational mathematical expressions and theoretical underpinnings for these methodologies are detailed in [28], [29], and are therefore not reiterated in this manuscript. The dynamic attributes of the system are meticulously evaluated through time-domain simulations, which involve solving a series of differential-algebraic equations at incremental time intervals. In addition, the system is subjected to linearisation at a specific operational point to facilitate the calculation of damping ratios, natural frequencies, eigenvalues, and eigenvectors for all modes, along with the derivation of the system's A, B, C, and D matrices.

The insights from modal examination highlight the occurrence of undamped oscillations spanning different areas within the system. As depicted in Fig. 11, the frequency plots from the simulation show a clear division where generators 1 and 2 in one region oscillate in unison against generators 3 and 4 in another region, evidenced by their nearly 180-degree phase difference, indicating opposite movements. This phenomenon is further substantiated through modal analysis.

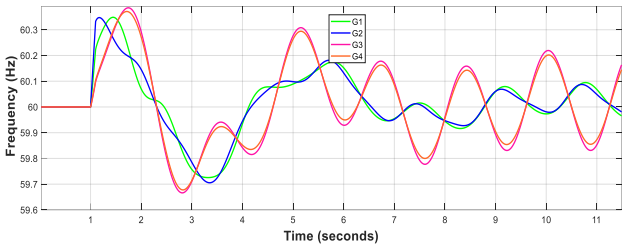


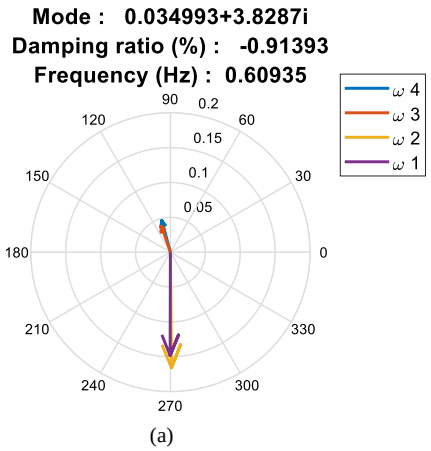
Fig. 11. Time-domain simulation of the system without controller.

Figure 11 shows the behaviour of the system in a time-domain simulation without the control mechanism. Table I lists the modes with minimal damping ratios, along with their associated frequencies and eigenvalues, pinpointing a mode with a negative damping ratio as a signifier of system instability. This specific mode, with a frequency of 0.60935 Hz, falls within the typical range for inter-area oscillation frequencies and is identified as a primary contributor to instability and oscillations across the system, implicating all rotating components along the interconnecting lines.

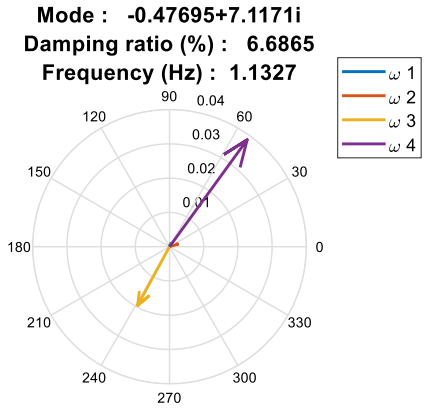
TABLE I. MODAL ANALYSIS RESULTS - WITHOUT CONTROLLER.

Damping ratio (%)	Frequency (Hz)	Mode	State_name
-0.0091393	0.60935	0.034993 + 3.8287i	'Si2q(3)', No.1
0.066865	1.1327	-0.47695 + 7.1171i	'Edp(4)', No.6
0.085341	1.0879	-0.58546 + 6.8352i	'Si2q(1)', No.7

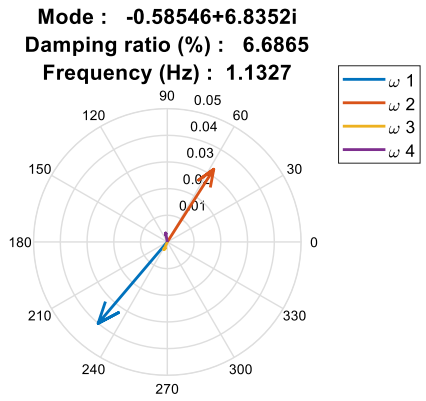
The mode shapes for the modes identified in Table I are graphically represented in Fig. 12, elucidating two distinct oscillatory behaviours within the system: inter-area and local oscillations between plants. The first mode shape illustrates the oppositional oscillation of generators across different areas, while the other two mode shapes depict oscillations within the same area, which have a lesser impact on overall stability and can be mitigated through localised power system stabilisers (PSSs). These mode shapes further confirm the dominance of Area 2 generators in influencing the inter-area oscillatory mode, a conclusion drawn from the vector lengths in Fig. 1 and oscillation amplitudes in Fig. 11.



(a)



(b)



(c)

Fig. 12. (a) Mode-shape plots of the inter-area mode ; (b) Local mode 1; (c) Local mode 2.

The system's pole-zero map, illustrated in Fig. 13, reveals an unstable mode characterised by a positive real component of the eigenvalue (0.034993 + 3.8287i).

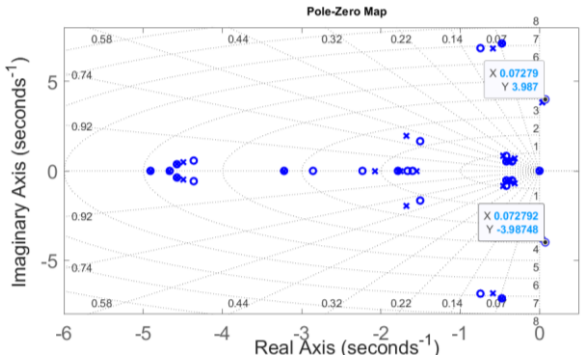
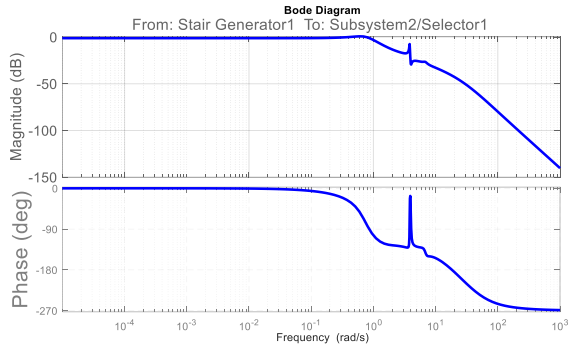


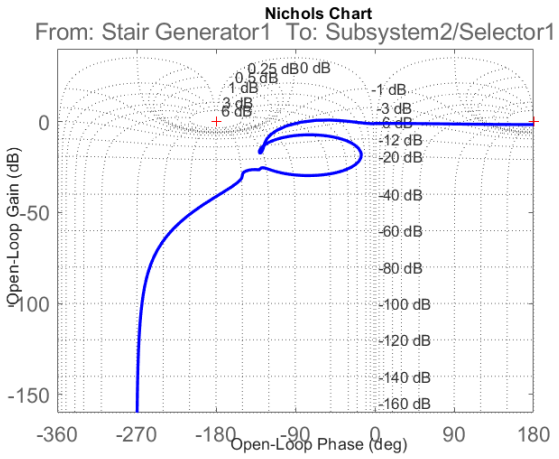
Fig. 13. Pole-zero map of the system without controller.

The results of the frequency response analysis, including the Bode, Nyquist, and Nichols plots, are presented in Fig. 14. These plots collectively indicate a pronounced shift in gain and phase, signaling potential instability within the system, in contrast to the smoother transitions observed in stable systems across a broad frequency spectrum. Specifically, the Nyquist plot underscores this instability through the encirclement of the critical point (-1, 0), while the Nichols plot corroborates this finding by highlighting the trajectory of the critical point at the coordinate (180 °, 0).

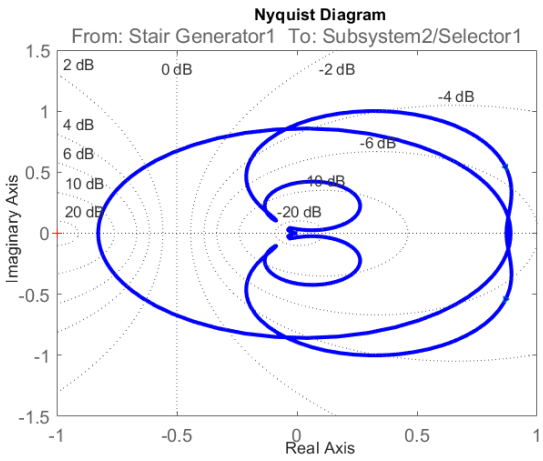
Figure 15 shows a participation factor map, derived from the right and left eigenvectors of the system post-linearisation. This map visually represents the influence of various state variables on the system's modes, particularly highlighting electromechanical variables such as rotor angle and machine speed. The map reveals that generators 3 and 4 in Area 2 have a pronounced impact on inter-area oscillatory modes, as indicated by the highlighted regions in the southwest quadrant of the map.



(a)



(b)



(c)

Fig. 14. Frequency response analysis: (a) Bode plot; (b) Nichols plot; (c) Nyquist plot.

The discussion thus far has focussed on the dynamics of the system in the absence of a control mechanism. The training process of the control mechanism and its efficacy are investigated, initiated by introducing a three-phase fault disturbance to induce significant oscillations within the system. The training results are depicted in Fig. 16, illustrating the episodes' reward, the average reward, and the Q-value trends of the episode. The training process starts with lower rewards, gradually increasing as the agent explores and exploits the environment, culminating when further improvements in reward become marginal.

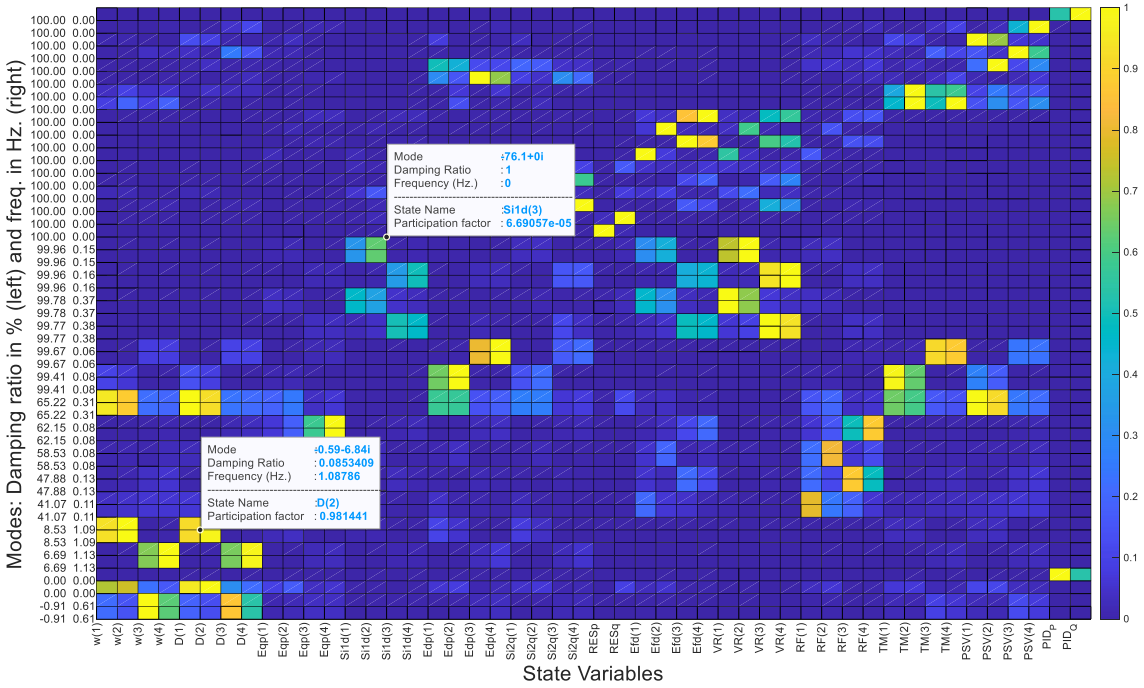


Fig. 15. Participation factor map of the linearised system.

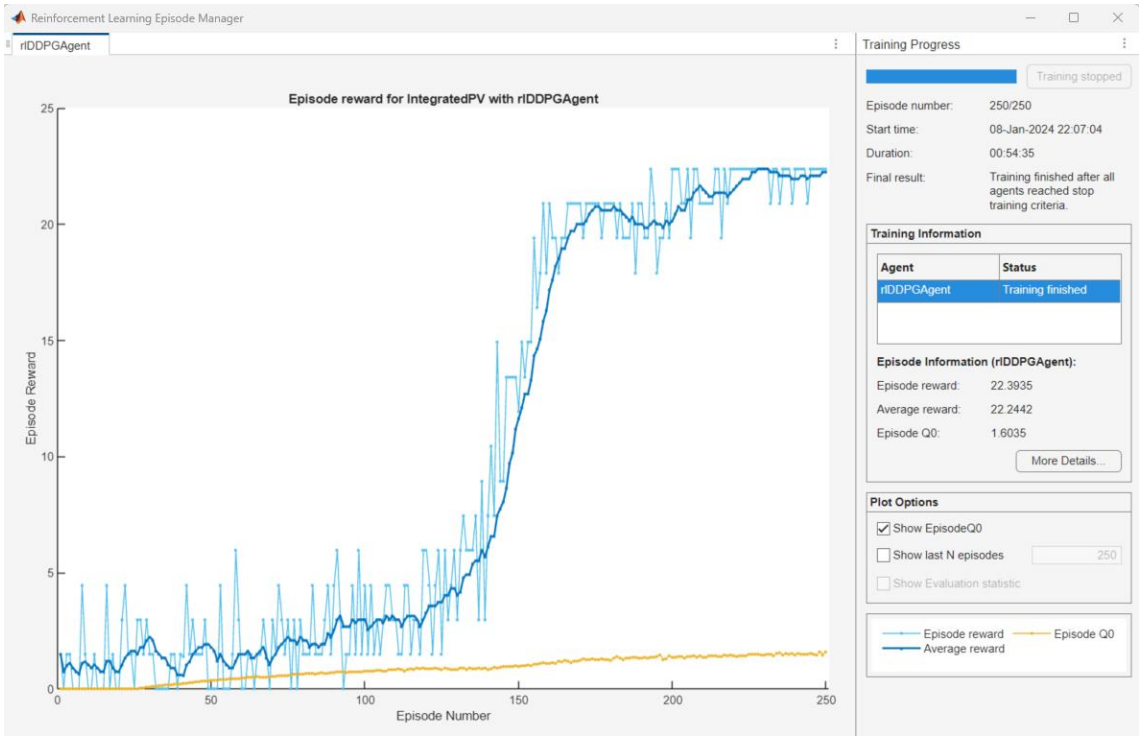


Fig. 16. Training episode reward as a function of episode number.

Post-training, the system equipped with the newly designed controller is analysed under realistic communication delay scenarios, both constant and variable, as shown in Fig. 17. Modal analysis post-controller implementation indicates improved damping of the inter-area mode, transitioning it to a stable state, with characteristics detailed in Table II. The controller efficacy is slightly better under constant delay conditions compared to variable delays, which is expected due to the potential extension of variable delays to 350 ms.

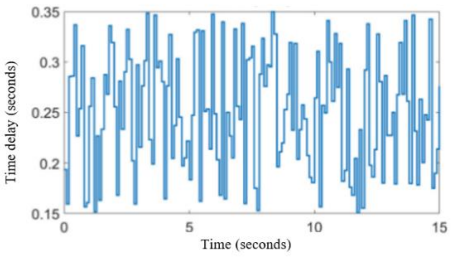
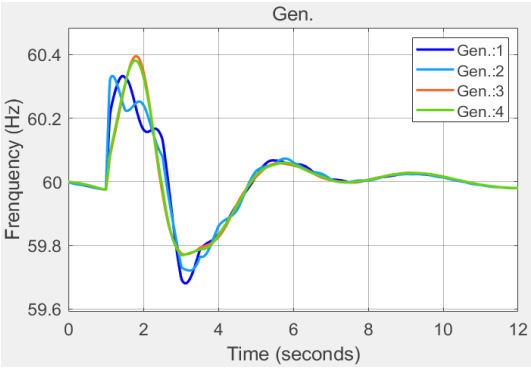


Fig. 17. Communication-delay variation applied to the controller input signal.

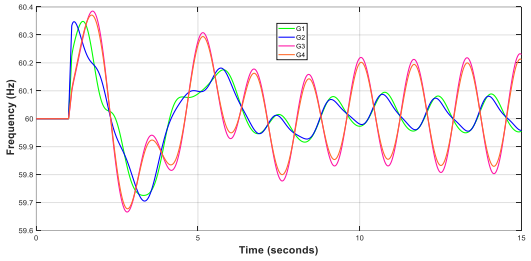
TABLE II. MODAL ANALYSIS RESULTS - WITH CONTROLLER AND DELAY.

Delay	Damping ratio (%)	Frequency (Hz)	Model
0.25 s	11.3	0.5730	-0.406 + 3.58i
0.15 s–0.35 s	9.63	0.6000	-0.36 + 3.75i

Figure 18 compares the system performance with and without the controller through time-domain simulations, highlighting the controller's role in damping oscillations. The frequency of the inter-area mode remains relatively constant across different scenarios, underscoring the controller's objective to dampen oscillations by shifting the real parts of the eigenvalues to the left on the complex plane, without altering the mode frequencies.



(a)



(b)

Fig. 18. (a) System analysis with controller time-domain simulation and (b) without controller.

The proposed control strategy, demonstrated through this study, holds potential for enhancing power system stability by damping inter-area oscillations, which, if inadequately damped, can lead to system instability. The control mechanism requires global information due to the oscillations resulting from the interaction between generators in different areas connected by weak tie-lines. Apart from damping inter-area oscillations, the control strategy can also mitigate local interactions within a single area, potentially eliminating the need for a PSS by incorporating auxiliary control elements in solar plants.

VII. CONCLUSIONS

This study presents a reinforcement learning-based controller, specifically designed for a solar facility connected to a fragile interconnection in a dual-region setup, with the goal of reducing cross-regional vibrations. Training of this agent is conducted using the deep deterministic policy gradient (DDPG) method. The reward mechanism of the controller is influenced by a discrete inverse function based on the error signal. An in-depth discussion is provided on the deep learning networks used for both the actor and the critic functions within the controller. The input to the controller is sourced from distant signals, obtained through phasor measurement units (PMUs) positioned at the generator locations, part of the broader area measurement infrastructure. The study conducts an extensive evaluation of the system performance with and without the implemented controller, employing various technical and analytical tools, including simulations in the time domain, analysis of frequency response, and maps of pole-zero and participation factors. The design also takes into account the potential impact of communication delays on oscillations to ensure the effectiveness of the controller under such conditions. MATLAB and Simulink were used to create a series of simulation tools, which encouraged further academic inquiry. The controller showed notable efficacy in quelling cross-regional oscillations, thus addressing the main issue of the study. The training phase was characterised by its efficiency and rapid convergence. It is important to note that the proposed control strategy is an unsupervised learning method that relies heavily on extensive data and computational resources for the training phase. When designing the controller, it is crucial to meticulously evaluate all possible variables, including system dynamics, noise, and perturbations. Should the policy prove to be less than optimal, the subsequent redesign, retraining, and evaluation processes could demand significant computational resources and time.

The results of the modal analysis for the system without a controller, presented in Table I, reveal various damping ratios and frequencies for different modes. The system shows a negative damping ratio of -0.0091393 % at a frequency of 0.60935 Hz for mode "*Si2q*(*3*)", No. 1, indicating instability. For mode "*Edp*(*4*)", No. 6, a damping ratio of 0.066865 % and a frequency of 1.1327 Hz are observed, while mode "*Si2q*(*1*)", No. 7, shows a damping ratio of 0.085341 % at 1.0879 Hz. When a controller with delay is introduced, as shown in Table II, there is a significant improvement in system stability. With a delay of 0.25 seconds, the damping ratio increases to 11.3 % with a frequency of 0.5730 Hz, and for a delay range of 0.15 to 0.35 seconds, the damping ratio is 9.63 % at a frequency of 0.6000 Hz. This enhancement indicates that the controller effectively stabilises the system, reducing the frequency slightly while significantly increasing the damping ratios.

CONFLICTS OF INTEREST

The author declares that he has no conflicts of interest.

REFERENCES

[1] Y. Hashmy, Z. Yu, D. Shi, and Y. Weng, "Wide-area measurement system-based low frequency oscillation damping control through reinforcement learning", *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5072–5083, 2020. DOI: 10.1109/TSG.2020.3008364.

[2] I. Abdulrahman, R. Belkacemi, and G. Radman, "Power oscillations damping using wide-area-based solar plant considering adaptive time-delay compensation", *Energy Syst.*, vol. 12, pp. 459–489, 2021. DOI: 10.1007/s12667-019-00350-2.

[3] D. Cao *et al.*, "Reinforcement learning and its applications in modern power and energy systems: A review", *J. Modern Power Syst. Clean Energy*, vol. 8, no. 6, pp. 1029–1042, 2020. DOI: 10.35833/MPCE.2020.000552.

[4] M. Sewak, *Deep Reinforcement Learning: Frontiers of Artificial Intelligence*. Springer Singapore, 2019. DOI: 10.1007/978-981-13-8285-7.

[5] H. Dong, Z. Ding, and S. Zhang, *Deep Reinforcement Learning:*

Fundamentals Research and Applications. Springer Singapore, 2020. DOI: 10.1007/978-981-15-4095-0.

[6] G. Zhang, W. Hu, J. Zhao, D. Cao, Z. Chen, and F. Blaabjerg, "A novel deep reinforcement learning enabled multi-band PSS for multi-mode oscillation control", *IEEE Trans. Power Syst.*, vol. 36, no. 4, pp. 3794–3797, 2021. DOI: 10.1109/TPWRS.2021.3067208.

[7] P. Gupta, A. Pal, and V. Vittal, "Coordinated wide-area damping control using deep neural networks and reinforcement learning", *IEEE Trans. Power Syst.*, vol. 37, no. 1, pp. 365–376, 2022. DOI: 10.1109/TPWRS.2021.3091940.

[8] H.-J. Lee, S.-S. Jhang, W.-K. Yu, and J.-H. Oh, "Artificial neural network control of battery energy storage system to damp-out inter-area oscillations in power systems", *Energies*, vol. 12, no. 17, p. 3372, 2019. DOI: 10.3390/en12173372.

[9] A. Younesi, H. Shayeghi, and M. Moradzadeh, "Application of reinforcement learning for generating optimal control signal to the IPFC for damping of low-frequency oscillations", *Int. Trans. Electr. Energ. Syst.*, vol. 28, p. e2488, 2017. DOI: 10.1002/etep.2488.

[10] F. Liu, Y. Li, M. Wu, Y. Zhou, and R. Yokoyama, "Robust wide-area damping controller design for inter-area oscillations with signals' delay", *IEEJ Trans. Electr. Electronic Eng.*, vol. 11, no. 2, pp. 206–215, 2016. DOI: 10.1002/tee.22208.

[11] C. Chen, M. Cui, F. Li, S. Yin, and X. Wang, "Model-free emergency frequency control based on reinforcement learning", *IEEE Trans. Industr. Inf.*, vol. 17, no. 4, pp. 2336–2346, 2021. DOI: 10.1109/TII.2020.3001095.

[12] G. Zhang *et al.*, "Deep reinforcement learning-based approach for proportional resonance power system stabilizer to prevent ultra-low-frequency oscillations", *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5260–5272, 2020. DOI: 10.1109/TSG.2020.2997790.

[13] Z. Yan, Y. Xu, Y. Wang, and X. Feng, "Deep reinforcement learning-based optimal data-driven control of battery energy storage for power system frequency support", *IET Gener. Transm. Distrib.*, vol. 14, no. 25, pp. 6071–6078, 2020. DOI: 10.1049/iet-gtd.2020.0884.

[14] M. Abouheaf, W. Gueaieb, and A. Sharaf, "Load frequency regulation for multi-area power system using integral reinforcement learning", *IET Gener. Transm. Distrib.*, vol. 13, no. 19, 2019. DOI: 10.1049/iet-gtd.2019.0218.

[15] Y. Zheng *et al.*, "Power system load frequency active disturbance rejection control via reinforcement learning-based memetic particle swarm optimization", *IEEE Access*, vol. 9, pp. 116194–116206, 2021. DOI: 10.1109/ACCESS.2021.3099904.

[16] Z. Yan and Y. Xu, "Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search", *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1653–1656, 2019. DOI: 10.1109/TPWRS.2018.2881359.

[17] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision", 2021. arXiv: 2102.01168.

[18] T. Yang, L. Zhao, W. Li, and A. Y. Zomaya, "Reinforcement learning in sustainable energy and electric systems: A survey", *Annual Reviews in Control*, vol. 49, pp. 145–163, 2020. DOI: 10.1016/j.arcontrol.2020.03.001.

[19] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview", *CSEE J. Power and Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2020. DOI: 10.17775/CSEEJPES.2019.00920.

[20] M. Glavic, "(Deep) Reinforcement learning for electric power system control and related problems: A short review and perspectives", *Annual Reviews in Control*, vol. 48, pp. 22–35, 2019. DOI: 10.1016/j.arcontrol.2019.09.008.

[21] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids", *CSEE J. Power and Energy Syst.*, vol. 4, no. 3, pp. 362–370, 2018. DOI: 10.17775/CSEEJPES.2018.00520.

[22] R. Siraskar, "Reinforcement learning for control of valves", *Mach. Learn. Appl.*, vol. 4, art. 100030, 2021. DOI: 10.1016/j.mlwa.2021.100030.

[23] I. Abdulrahman and G. Radman, "Wide-area-based adaptive neuro-fuzzy SVC controller for damping interarea oscillations", *Can. J. Electric. Comput. Eng.*, vol. 41, no. 3, pp. 133–144, 2018. DOI: 10.1109/CJECE.2018.2868754.

[24] J. H. Chow, P. W. Sauer, M. A. Pai, *Power System Dynamics and Stability*, 2nd ed., Wiley-IEEE Press, 2017.

[25] A. Ellis, M. R. Behnke, and R. T. Elliott, "Generic solar photovoltaic system dynamic simulation model specification", Office of Scientific and Technical Information (OSTI), Oct. 2013. DOI: 10.2172/1177082.

[26] M. R. Islam, F. Rahman, and W. Xu, *Advances in Solar Photovoltaic Power Plants*. Springer Berlin, Heidelberg, 2016. DOI: 10.1007/978-3-662-50521-2.

[27] WECC Renewable Energy Modeling Task Force (2014) WECC solar plant dynamic modeling guidelines. [Online]. Available: https://www.wecc.org/Reliability/WECC%20Solar%20Plant%20Dynamic%20Modeling%20Guidelines.pdf

[28] P. Kundur *et al.*, "Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions", *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1387–1401, 2004. DOI: 10.1109/TPWRS.2004.825981.

[29] B. Pal and B. Chaudhuri, *Robust Control in Power Systems*. Springer New York, NY, 2005. DOI: 10.1007/b136490.