Scientific Reports

https://doi.org/10.1038/s41598-026-55262-w

Article in Press

Double deep reinforcement learning twin-delayed agents for performance improvement of a grid-connected wave energy conversion system

Received: 18 March 2026

Accepted: 22 May 2026

Published online: 27 May 2026

Cite this article as: Aldawsari F., Mahdy A., Ali Z.M. *et al.* Double deep reinforcement learning twin-delayed agents for performance improvement of a grid-connected wave energy conversion system. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-55262-w

Faisal Aldawsari, Ahmed Mahdy, Ziad M. Ali, Ahmed F. Zobaa, Shady H. E. Abdel Aleem & Hany M. Hasanien

Double Deep Reinforcement Learning Twin-Delayed Agents for Performance Improvement of a Grid-Connected Wave Energy Conversion System

Faisal Aldawsari [1], Ahmed Mahdy [2], Ziad M. Ali [3], Ahmed F. Zobaa [4,*] , Shady H. E. Abdel Aleem [5], and Hany M. Hasanien [6]

1. Electrical Engineering Department, College of Engineering at Wadi Addawaser, Prince Sattam bin Abdulaziz University, Wadi Addawaser 11991, Saudi Arabia; f.aldawsari@psau.edu.sa
2. Electrical Power and Machines Department, Faculty of Engineering, Ain Shams University, Cairo, 11517, Egypt; ahmedmahdy@khadijaacademy.com
3. Electrical Engineering Department, College of Engineering at Wadi Addawaser, Prince Sattam bin Abdulaziz University, Wadi Addawaser 11991, Saudi Arabia; dr.ziad.elhalwany@aswu.edu.eg
4. Department of Engineering; College of Engineering, Design and Physical Sciences; Brunel University of London; Uxbridge UB8 3PH; United Kingdom, azobaa@ieee.org
5. Basic Sciences Council, Academy of Scientific Research and Technology, Cairo, 11516, Egypt; engyshady@ieee.org
6. Department of Electrical Engineering, College of Engineering, University of Sharjah, Sharjah, 27272, United Arab Emirates; hanyhasanien@ieee.org

*Corresponding author: Ahmed F. Zobaa (azobaa@ieee.org)

Abstract: This study introduces a novel approach using two deep learning agents, trained with the twin-delayed deep deterministic policy gradient (TD3) algorithm, to replace the PI controllers used for the control of grid-connected Archimedes Wave Swing (AWS) wave energy conversion systems. The generator converter's controller has two mandatory objectives: minimizing losses in the stator and maximizing energy extraction from incident sea waves. These goals are achieved by controlling the generator's *dq* currents using a TD3 agent on the rectifier side. In addition, the grid-side inverter's controller is responsible for regulating both the DC link and the point-of-common-coupling voltages. In the new configuration, two approaches are proposed in this work: either a single deep learning agent replaces the four proportional-integral (PI) controllers on the inverter side, or a hybrid approach combining two PI controllers with a TD3 agent. To verify the reliability of the TD3

agents, the system is analyzed in both steady and transient states under fault conditions. Furthermore, the TD3 agents' performance is benchmarked against the classical PI controller configuration in MATLAB Simulink. The results demonstrate better dynamic and steady-state responses from the hybrid-TD3 agent on the grid side than from the full PI classical configuration.

Keywords: Archimedes wave swing; deep learning; power system control; twin-delayed deep deterministic policy gradient; wave energy conversion systems.

1. Introduction

Ocean wave energy, a resource with immense global potential, reaching up to 32,000 TWh/yr, has attracted considerable interest recently [1]. Unlike solar and wind, wave energy offers greater predictability and availability. This makes it a strong candidate for integration into modern power systems. By 2035, the levelized cost of electricity generated from good wave energy resources could drop under 70 €/MWh [2]. However, wave energy converters (WECs) have several drawbacks, including maximizing energy extraction from waves, economic feasibility (high costs), and grid integration. This work focuses on addressing the problem of successful grid connection and surviving grid faults. In the literature, numerous wave energy converters, including the AWS, oscillating water column (OWC), Wave Dragon, and Pelamis, have been controlled and optimized to maximize energy extraction and ensure strong grid connection. For instance, a wave-to-grid (W2G) architecture has been developed for a point absorber wave energy converter to enhance energy extraction and provide a smooth power supply to the grid [3]. The mechanical variables are estimated using an extended Kalman filter, enabling sensorless control of the device. Moreover, these variables are employed as input to the model-predictive control (MPC) technique for controlling the generator side. In addition, a battery energy storage system is used to provide smooth power to the power grid. Another study addressed the problem of peak-to-average power in wave energy converters [4]. Strong, instantaneous sea waves can cause an extreme burst of power (peak) for a short period. This requires an oversized generator and turbine to handle this power for a short time, leading to a higher cost of the wave energy converter. The authors deploy a high-speed stop valve in series with the OWC air turbine to restrict air flow during periods of high power. The results confirmed the

viability of the control method and achieved a low peak-to-average power ratio. However, the disadvantage is the violation of the constraints for a very short period. Furthermore, a study focuses on maximizing energy extraction from AWS devices by regulating the *q*-axis current of the generator-side to control the generator velocity [5]. The system uses lithium-ion batteries as an energy storage element to exchange reactive power with the AWS device as needed, thereby eliminating the need for the grid to reverse power. Moreover, a study proposes a W2G architecture for a floating wave energy converter that uses a LiTe-Con controller to enhance wave energy extraction [6]. An ultra-capacitor is integrated into the DC link to provide the reactive power required by the wave energy converter. Both generator-side and grid-side converters employ Lyapunov-based nonlinear controllers to minimize copper losses and current errors, achieve a unity power factor, and regulate the DC link voltage. Additionally, recent research has focused on applying advanced optimization algorithms and exploring new energy vectors to enhance the performance of renewable energy systems. For example, the Red-Tailed Hawk optimization algorithm has been applied for optimal PV array reconfiguration under partial shading conditions, significantly improving energy yield in practical applications such as irrigation systems [7]. Furthermore, the potential of green hydrogen as a complementary energy carrier for storing surplus renewable energy has been extensively investigated, with studies evaluating regional opportunities and challenges for large-scale hydrogen production using solar resources [8]. Although these studies focus on photovoltaic systems and hydrogen production, their core concepts are highly relevant to wave energy systems. In the context of the AWS devices, the PI controllers were tuned using the Coot optimization algorithm to improve the dynamic stability of the grid-connected AWS system under severe fault conditions [9]. Furthermore, recent advances in control theory have enabled the successful application of fractional-order and nonlinear PID-based controllers across various industrial domains. For instance, a study in [10] demonstrated the effectiveness of a multi-stage FOPI controller for nonlinear motor systems, while [11] proposed a hyperbolic tangent-based PID for pressure control. Although these studies focus on different applications, they highlight the growing interest in developing robust controllers for nonlinear, uncertain systems like the AWS converter. For instance, a study focused on adopting the fractional-order PID (FOPID) for controlling the AWS system [12]. The FOPID controllers are tuned using

the hybrid jellyfish search and particle swarm optimization, achieving higher performance than PI controllers. In [13], a hydraulic energy-storage wave energy converter temporarily stores energy harvested from waves in high-pressure accumulators. At a certain pressure, the converter starts the generation stage and provides power to the grid. When the pressure goes down below a certain threshold, the generator stops and returns to the energy storage phase. The system incorporates batteries to further smooth the power given to the grid. In [14], five PI controllers are employed to manage the rectifier and inverter in an AWS converter. In addition to the coupling of a distributed battery energy storage (DBES) with a DC/DC converter integrated to the DC link to smooth the power provided to the electrical grid. The same control configuration can be found in [15], where a supercapacitor energy storage is employed instead of the DBES to smooth the output power. Furthermore, a study proposes a hybrid energy storage system comprising a supercapacitor for short-term, fast response and batteries for long-term, slow response to smooth the injected power from a wave energy park composed of multiple wave energy converters [16]. Moreover, a study proposes a low-voltage ride-through controller that increases reactive power injection during fault conditions for AWS systems [17]. The magnitude of the reactive power injection is dependent on the dip in the voltage at the point of common coupling V_{PCC}. Moreover, the generator's active power is set to zero, and a braking chopper is used to prevent overvoltage in the DC link.

The primary focus of this paper is on the AWS, as it is a widely used wave energy converter [18]. In this study, the main objective is to implement a novel control methodology that hasn't been explored within the context of AWS technology. Recent AI-driven developments have opened new avenues for revolutionary control methods in renewable energy systems. Deep learning agents demonstrate strong capabilities for handling nonlinear, time-varying, and uncertain dynamics that align with the operational environment of wave energy converters. Moreover, deep learning approaches can directly learn optimal control policies by acquiring data through actions and rewards received during interaction with the environment. There are various deep learning algorithms for training an agent, including deep Q-learning (DQN), soft actor-critic (SAC), proximal policy optimization (PPO), twin-delayed deep deterministic policy gradient (TD3), and deep deterministic policy

gradient (DDPG), among others. For instance, an RL-beta method is adopted for maximum power point tracking (MPPT) in a solar energy system [19]. Beta is calculated from voltage and current measurements, which the agent uses to produce a discrete action: a change in the duty cycle. Moreover, a study uses the TD3 algorithm to train an agent controlling a multi-level inverter in a solar energy system [20]. The TD3 agent is employed to minimize the errors in the *dq* currents by selecting the optimal *dq* voltages of a solar inverter. Another work uses the DQN to train an agent to directly select the optimal switching action for a three-level inverter [21]. The results are compared with the conventional MPC strategy. In addition, a study uses the dandelion optimizer with either DDPG or PPO for MPPT in a 100 MW PV system [22]. The dandelion optimizer provides the reference voltage for MPPT. The PPO or DDPG is used to minimize the error between the reference and the actual system voltages by selecting the optimal duty cycle for pulse-width modulation (PWM) of a boost converter in the PV system. Furthermore, a study proposes a physics-guided deep learning framework in place of conventional lookup tables for real-time wind farm flow control under time-varying conditions [23]. This approach utilizes a clustering-based precomputation strategy that cuts computational overhead by 85.5% while achieving 40% better control performance than standard methods. In addition, a paper focuses on solving the voltage fluctuations in a multi-feeder distribution system using deep learning agents [24]. The system configuration consists of a main agent responsible for controlling the tap position of the on-load tap changer and sub-agents for adjusting the reactive power of each inverter in the associated PV systems. The results indicate a voltage deviation of only 1.28% from the optimum. Moreover, a deep learning framework combined with MPC is proposed for energy management in a building [25]. The MPC employs an approximate model of the system to filter and ensure that the agent's action remains within the constraints. For wave energy converters, the researchers focus on maximizing energy extraction from waves. For instance, DQN is employed for a point absorber WEC [26]. The agent takes an action by changing the load resistance to maximize energy extraction. Furthermore, DDPG was applied in AWS devices to maximize energy extraction by changing the reference quadrature-axis current ($i_{q\text{-}ref}$) to vary the generator force under changing sea states [27]. This controller was compared with multiple energy maximization strategies, and the

results confirm its greater energy extraction capability. In addition, many studies use the SAC algorithm to control the power-take-off (PTO) force to increase energy harvest [28], [29], [30]. In [31], the oscillating wave surge converter's energy extraction is maximized by the formulation of an adaptive damping policy under different sea conditions. Several algorithms, including TD3, SAC, and PPO, are employed to achieve this objective.

In this work, the main goal is to replace the classical PI controller configuration with two deep learning agents. These agents minimize the errors in control signals by providing suitable reference *dq* voltages, as explained later. These voltages can take any value between -1 and 1. This leads to the requirement of a continuous action-type algorithm. In this study, the TD3 is employed due to its numerous advantages over DQN and DDPG. The novel contributions of this work are summarized as follows:

- The TD3 deep learning agent is trained for 10 seconds to minimize the generator losses and maximize energy extraction by controlling the generator's *dq* currents. Subsequently, the agent is evaluated for an additional 40 s (unseen states) to verify its reliability.
- The generator-side TD3 agent was compared with the classical configuration of PI controllers and achieved lower ISE for the *dq* currents. Moreover, the agent was subjected to various sea states and to changes in the floater mass to assess its reliability.
- Additionally, another two TD3 agents for the grid-side are trained for 10 s (one under steady-state and the other under transient state).
- The two grid-side agents and the PI controllers are benchmarked against each other under the steady-state for 50 s. In addition, the three configurations are analyzed during the transient state under different grid faults, including three-line (LLL), double-line to ground (2LG), line to ground (LG), three-line to ground (3LG), and line-line (LL) faults, to assess the efficacy of the agents. These results were obtained by simulating the grid-connected AWS system in Simulink.
- The controller efficacy was analyzed under the effect of different grid SCR values (1.5, 3, 5). The controller demonstrated its efficiency in the steady state and during transient operation (SCR = 5), further emphasizing its superior performance.

The remaining sections of this paper are organized as follows: the AWS system model and the linear generator are discussed in Section 2. The classical PI control loop and the new control strategy using double deep learning agents are detailed in Section 3. Section 4 presents benchmarking of the double-agent configuration against the classical configuration in steady and transient states. The concluding remarks are summarized in Section 5.

2. Modeling of the AWS system

2.1. Modeling of the AWS device

The AWS is a submerged WEC that relies on the reciprocating motion of its floater (up-and-down movement) caused by sea waves. A linear generator converts the AWS motion to usable electrical power. Newton's second law formulates the equations of motion presented by Eqs. (1) and (2) [32]:

$$m_f\frac{dv}{dt} = F_{drag}+F_{grav}+F_{hS}+F_{rad}+F_{sp}+F_{wb}+F_{gen}+F_{end}+F_e+F_{bear} \quad (1)$$

$$v = \frac{dx}{dt} \quad (2)$$

x and v are the position and velocity of the AWS floater, respectively. When F_e is positive, the floater moves downwards, which corresponds to the positive direction of x and v. The resistance encountered by the floater during its oscillatory movement is represented by the drag force (F_{drag}). This force is presented in Eq. (3) and has two drag coefficients (C_{DUP}) and (C_{DDW}) for both upward and downward motions. In addition, water density and the floater's outer surface area are symbolized by ρ and S_F. The gravitational force (F_{grav}) is articulated by Eq. (4). In this equation, m_f and g are the floater's mass and the gravitational acceleration.

$$F_{drag} = \begin{cases} -\frac{1}{2}\rho S_F v|v|C_{DUP}, & v \geq 0 \\ -\frac{1}{2}\rho S_F v|v|C_{DDW}, & v < 0 \end{cases} \quad (3)$$

$$F_{grav} = -m_f g \quad (4)$$

The difference in pressure on the buoy causes the hydrostatic force (F_{hS}). F_{hS} is presented in Eq. (5) and consists of multiple parameters, including the tide's level (η), the atmospheric pressure (p_{amp}), the floater height (h_f), the inner area of the floater (S_f), and the depth at mid-position (d_f) [33].

$$F_{hS} = -S_F(\rho g(d_f + \eta - x) + p_{amp}) + (S_F - S_f)(\rho g(d_f + \eta + h_f - x) + p_{amp}) \quad (5)$$

The radiation of water waves is caused by the oscillatory motion of the floater. This phenomenon is expressed by the radiation force (F_{rad}). F_{rad} is presented by Eq. (6). In this equation, the added water mass and the radiation damping kernel are denoted by $R(t)$ and m_{add}.

$$F_{rad} = -m_{add}\frac{dv}{dt} - \int_0^t R(t-\tau)v(\tau)d\tau \quad (6)$$

A spring is used to restore the AWS floater to the equilibrium position. The effect of spring is presented by the spring force (F_{sp}) expressed by Eq. (7). F_{sp} is dependent on the length at equilibrium (L_{eq}), spring force in the equilibrium state ($F_{sp\text{-}eq}$), the heat capacity rate (γ), and the floater's equilibrium position (x_{eq}).

$$F_{sp} = F_{sp\text{-}eq}\left(\frac{L_{eq}}{L_{eq}+x-x_{eq}}\right)^{\gamma} \quad (7)$$

The floater contains water brakes (F_{wb}) with a damping coefficient (β_{wb}) that prevent the floater from exceeding $x = 4$ m. In addition to the mechanical end stops provided to protect the floater from damage at $x = 4.5$ m. The force F_{wb} and F_{end} are formulated using Eqs. (8) and (9).

$$F_{wb} = -\beta_{wb}v|v|, \qquad x \geq 4 \text{ or } x \leq -4 \quad (8)$$

$$F_{end} = -\frac{v(m_{add} + m_f)}{0.1}, \qquad x \geq 4.5 \quad (9)$$

The excitation force of waves that results in the floater motion is symbolized by F_e. This force and its corresponding components are given in Eqs. (10)-(15):

$$F_e = \rho g S_F \eta(t) K_{pw} \quad (10)$$

$$\eta(t) = \sum_{i=1}^{N}\frac{H_i}{2}\sin(\omega_i t + \theta_i) \quad (11)$$

$$A_i = \frac{H_i}{2} = \sqrt{2S(\omega_i)\Delta\omega_i} \quad (12)$$

$$S(\omega) = \frac{487\,H_s^2}{T_p^4\omega^5}e^{-\frac{1948.2}{T_p^4\omega^4}} \quad (13)$$

$$K_{pw} = \frac{\cosh(k(h-d))}{\cosh kh} \quad (14)$$

$$\omega^2 = gk\tanh(kh) \quad (15)$$

F_e of irregular sea waves is produced by combining several sinusoidal regular waves (21 waves are selected with angular frequencies between 0.5 and 2.5 rad/s). The irregular waves' total elevation is denoted by ($\eta(t)$). A_i, ω_i, and θ_i represent the amplitude, the angular frequency, and the phase shift of each sinusoidal wave. The Bretschneider spectrum ($S(\omega)$) is adopted to form irregular sea waves, and the angular frequency difference between these waves is denoted by $\Delta\omega_i$ with a value of 0.1 rad/sec [34]. This spectrum relies on two main parameters, the significant height (H_s) and the peak energy period (T_p). Another important parameter is the wave pressure decay factor (K_{pw}). The pressure of sea waves is the highest at the surface. However, as depth increases, the effect of this pressure decreases. Hence, K_p is required to provide the actual pressure on the floater based on the depth (d), the wave number (k), and the distance to the seabed from the floater (h). Moreover, the wave particles have horizontal wave velocity (u) and acceleration ($\dot{u}$) that affects the AWS floater in the form of the horizontal force (F_H). F_H contributes to the frictional force (F_{bear}) that affects the AWS bearings with a friction coefficient (μ). F_H is dependent on the drag coefficient (c_D), the inertia coefficient (c_M), and the outer diameter (d_{out}). Finally, F_H is obtained by performing an integration of the wave particles' horizontal effect across the floater length. The equations are provided in Eqs. (16)-(21) [35].

$$F_{bear} = -\mu \cdot \text{sign}(v)\,|F_H| \tag{16}$$

$$dF_H(z,t) = c_M \rho \frac{\pi}{4} d_{out}^2 \dot{u}(z,t)dz + c_D \rho \frac{1}{2} d_{out} u(z,t)|u(z,t)|dz \tag{17}$$

$$F_H = \int_{z=0}^{z=h_f} dF_H(z,t) \tag{18}$$

$$u = \frac{H\omega}{2} \frac{\cosh(k(h-d))}{\sinh kh} \cos(\omega t) \tag{19}$$

$$\dot{u} = -\frac{H\omega^2}{2} \frac{\cosh(k(h-d))}{\sinh kh} \sin(\omega t) \tag{20}$$

$$d = d_f + h_f - z \tag{21}$$

2.2. Modeling of the linear generator

The AWS linear generator can be modeled using the conventional *abc* frame or the *dq0* frame. In this work, the *dq0* equations are used to model the generator. In [36], Feng proposes the equations that represent the generated terminal *dq* voltages (v_d and v_q) and currents (i_d and i_q) during both positive and negative floater

velocities. These formulas are essential for controlling the generator. These voltages are expressed by Eqs. (22) and (23):

$$v_d = -Ri_d + X_s i_q - L_s\left(\frac{\omega_{gen}}{|\omega_{gen}|}\right)\left(\frac{di_d}{dt}\right) \quad (22)$$

$$v_q = -Ri_q - X_s i_d - L_s\left(\frac{\omega_{gen}}{|\omega_{gen}|}\right)\left(\frac{di_q}{dt}\right) + \omega_{gen}\psi_{PM} \quad (23)$$

R and L_s symbolize the generator phase resistance and inductance. X_s denotes the phase reactance ($X_s = |\omega_{gen}|L_s$). ω_{gen} is the generator angular speed ($2\pi v/\lambda$). λ is the pole width. ψ_{PM} represents the permanent magnets' flux linkage. The generator output power (P_{gen}) and the force (F_{gen}) are linked together using Eq. (24).

$$F_{gen} = \frac{-P_{gen}}{v} = \frac{-3\omega_{gen} i_q \psi_{PM}}{2v} = \frac{-3\pi i_q \psi_{PM}}{\lambda} \quad (24)$$

The model parameters adopted for the AWS system are presented in Table 1.

Table 1

Numerical values of the AWS system parameters

Symbol	Value	Symbol	Value	Symbol	Value
m_f	4×10^5 kg	γ	1.4	ψ_{PM}	23 Wb
m_{add}	3.55×10^5 kg	S_F	95 m^2	λ	0.1 m
ρ	1025 kg/m^3	σ	4 m	C_{DDW}	0.4
p_{amp}	1×10^5 N/m^2	μ	0.1	C_{DUP}	0.2
β_{wb}	1.5×10^6 kg/m	η	0	R	0.29 Ω
N	21 waves	c_M	2	L_s	0.031 H
d_f	11 m	S_f	79 m^2	h_f	28.5 m
c_D	1	g	9.8 m/s^2	d_{out}	11 m

3. Control of an AWS WEC connected to the grid

The linear generator outputs a three-phase voltage characterized by fluctuating magnitude and frequency. This voltage can't be provided directly to the power grid without some modifications first. Therefore, a combined configuration of a rectifier, a DC link, and an inverter is adopted to improve suitability. The inverter is followed

by a simple RL filter to eliminate a significant portion of switching harmonics. The filter output is fed to a step-up transformer to raise the voltage to 25 kV, suitable for the medium voltage distribution network. The transformer is connected to the grid via double transmission lines. Fig. 1 depicts the system's architecture.

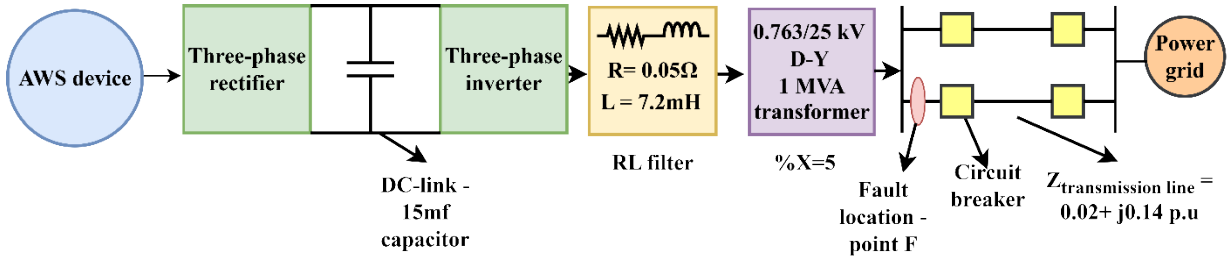


Fig. 1. The overall grid-connected architecture of the AWS device.

3.1. The mathematical formulation of the twin-delayed deep deterministic policy gradient algorithm (TD3)

TD3 is considered the upgraded version of the DDPG algorithm [37]. Both are used for continuous action spaces and employ an actor-critic architecture, which is appropriate for the control system. The classical DDPG algorithm uses one actor and one critic:

- The actor is the final neural network that provides the optimum action based on current observations.
- The critic estimates the quality of the action (Q-value) taken by the actor during the training process and hence controls the update procedure of the actor's neural network weights and biases.

The common problem with the DDPG's critic is the overestimation of the Q-value, or in other words, the critic's problem of being too optimistic, and the training instability. These problems are addressed in the TD3 as follows:

- Regarding the first problem, the TD3 adopts two critics (twin) instead of just one. During target value computation, the minimum of the two critics' values is used (clipped double Q-learning).
- With respect to the second issue, in DDPG, the actor is updated every step. However, TD3 introduces a delay in the update process (updates the actor once every two steps), which makes the training more stable.

- DDPG can have unstable updates due to errors in *Q*-values. Therefore, TD3 adds noise to the target action to make the policy more robust. It smooths the *Q*-value estimate around the target action.

The training process starts with the random initialization of the weights and biases of the actor (π_θ) and the two critics ($Q_{\phi 1}$ and $Q_{\phi 2}$) neural networks. In addition to copying the neural networks of the actor and the two critics to form the target networks $\pi_{\theta'}$, $Q_{\phi 1'}$, and $Q_{\phi 2'}$. π and Q signify the actor policy and the *Q*-value provided by the critic network, respectively. Moreover, θ, ϕ, θ', and ϕ' are the parameters of the actor, critic, target actor, and target critic networks, respectively. The actor's current policy takes an action with noise ($\epsilon_{explore}$) added for the exploration process during environment interaction presented by Eq. (25):

$$a = \pi_\theta(s) + \epsilon_{explore} \tag{25}$$

where a, s, and ϵ denote the action taken, current state, and the exploration noise. After a, the agent obtains a reward (r) and goes to a new state (s'). This information (s, a, r, and s') is collected in the experience replay buffer (D) for updating the neural networks. The agent keeps taking actions and collecting experiences until the replay buffer reaches its minimum size (a mini-batch). To train the neural networks, a random mini-batch is taken from the replay buffer. Then, the target action (a') is computed with noise (ϵ_{smooth}) for smoothing the *Q*-function using Eq. (26), and the target *Q*-value (y) is computed via Eq. (27):

$$a' = \pi_{\theta'}(s') + \epsilon_{smooth} \tag{26}$$

$$y = r + (1 - d)\gamma \min_{i = 1,2} Q_{\phi_i'}(s', a') \tag{27}$$

In Eq. (27), γ denotes the discount factor, which represents the weight provided for the future rewards and how much we care about those rewards in comparison with the instant reward from taking an action, and d presents the termination flag. The next step is updating the twin critic networks using the mean squared error between the predicted *Q*-value ($Q_{\phi_i}(s,a)$) and the target value (y). This is given by the loss function ($L(\phi_i)$) in Eq. (28):

$$L(\phi_i) = E_{(s,a,r,s',d)\sim D}\left[(Q_{\phi_i}(s,a) - y)^2\right] \quad (28)$$

where E is the expectation (average over one batch). To update the critic network, a gradient descent for $L(\phi_i)$ with respect to the network parameters (ϕ_i) is performed in Eq. (29):

$$L\nabla_{\phi_i} L(\phi_i) = E[2(Q_{\phi_i}(s,a) - y)\nabla_{\phi_i} Q_{\phi_i}(s,a)] \quad (29)$$

Update of the network parameters is achieved using a learning rate (α_Q) in Eq. (30):

$$\phi_i \leftarrow \phi_i - \alpha_Q \nabla_{\phi_i} L(\phi_i) \quad (30)$$

The actor is updated every couple of steps using the policy gradient in Eq. (31). In this equation, the weights are updated in the direction that maximizes the expected reward by the critic (gradient ascent). TD3 uses only one critic.

$$\nabla_\theta J(\theta) = E_{s\sim D}\left[\nabla_a Q_{\phi_1}(s,a)|_{a=\pi_\theta(s)} \nabla_\theta \pi_\theta(s)\right] \quad (31)$$

Like the critic, the actor network parameters are updated using a learning rate (α_a) as given in Eq. (32):

$$\theta \leftarrow \theta + \alpha_a \nabla_\theta J(\theta) \quad (32)$$

The critic and actor target networks are updated slowly every two steps using Eqs. (33) and (34):

$$\phi_i' \leftarrow \tau\phi_i + (1 - \tau)\phi_i' \quad (33)$$

$$\theta' \leftarrow \tau\theta + (1 - \tau)\theta' \quad (34)$$

where τ is the target network update rate. This completes the overall explanation of TD3. In the next part, the application of the TD3 agent to the rectifier and the inverter is discussed.

3.2. Control of the three-phase rectifier using PI controllers

The generator-side two-level rectifier comprises six IGBT switches. This converter's controller has two objectives. The minimization of losses by controlling i_d and maintaining it at zero ($i_{d\text{-ref}} = 0$), and the maximization of energy extraction

by controlling i_q at a certain value dependent on v. A simple technique, approximate current control, is adopted. The reference current ($i_{q\text{-}ref}$) is given in Eq. (35) [38]:

$$I_{q\text{-}ref} = 138v \tag{35}$$

The 138 value was selected by trial and error [38]. The control loop of these two currents is designed based on Eqs. (22) and (23). The control loop is formed of two PI controllers, to which the *dq* currents' errors (e_{id} and e_{iq}) are provided. The PI controllers generate the appropriate *dq* reference voltages. These voltages are converted into *abc* voltages ($V_{abc\text{-}ref}$) using the inverse Clarke-Park transformation. The transformation angle (θ_{t1}) is calculated using ($\frac{2\pi x}{\lambda} - \frac{\pi}{2}$). Finally, the switching actions are provided to the IGBTs by comparing $V_{abc\text{-}ref}$ with a 5-kHz triangular waveform. The control loop is depicted in Fig. 2.

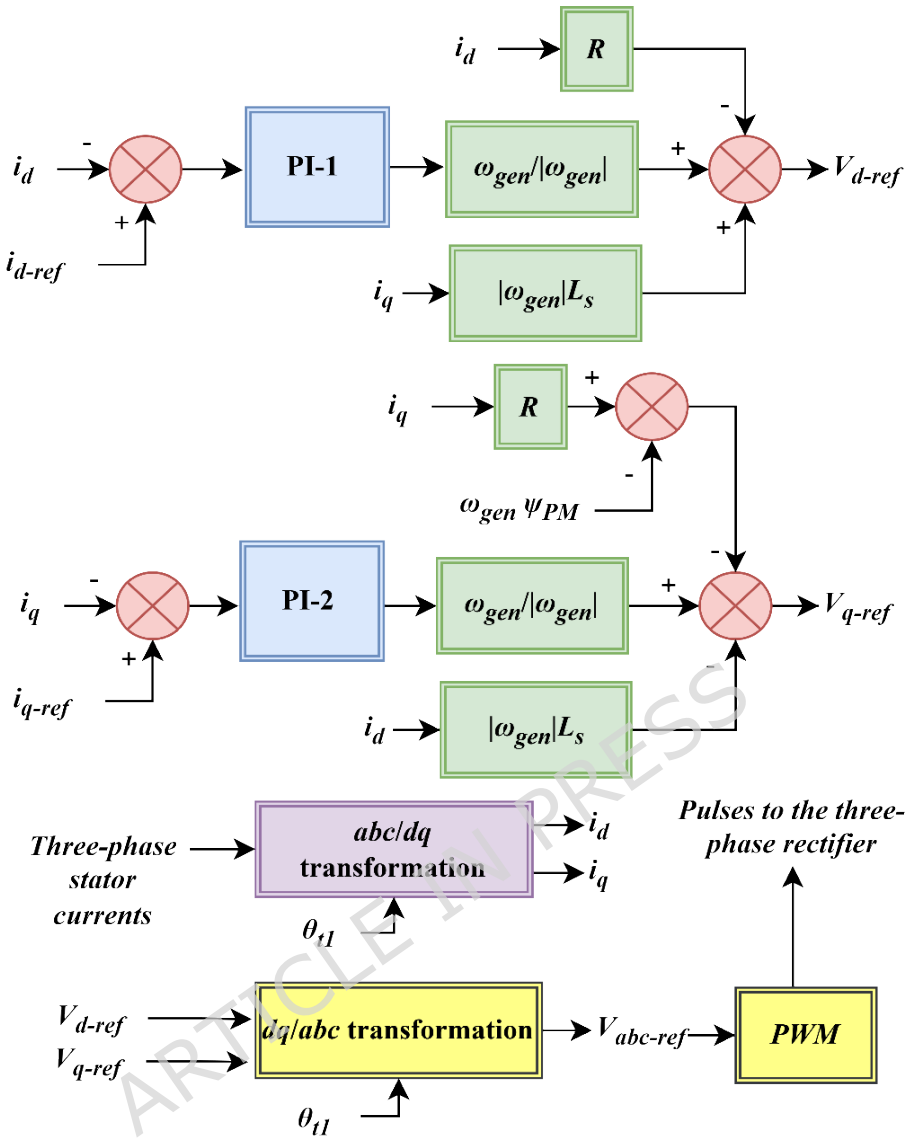


Fig. 2. Control loop of the generator side using PI controllers.

3.3. Control of the three-phase rectifier using a TD3 agent

The same objectives for the *dq* currents are accomplished via the generator-side TD3 agent. This agent provides the reference *dq* voltages based on some appropriate observations. The selected inputs upon which the agent takes decisions are the *dq* currents' errors (e_{id} and e_{iq}), the values of the *dq* currents, and v (as the direction of motion affects the control action). Subsequently, the agent would provide the reference *dq* voltages that are converted into *abc* voltages ($V_{abc\text{-}ref}$) using the inverse Clarke-Park transformation. The transformation angle (θ_{t1}) is calculated using ($\frac{2\pi x}{\lambda}$

$-\frac{\pi}{2}$). Finally, the switching actions are provided to the IGBTs by comparing $V_{abc\text{-}ref}$ with a 5-kHz triangular waveform. The control loop is depicted in Fig. 3.

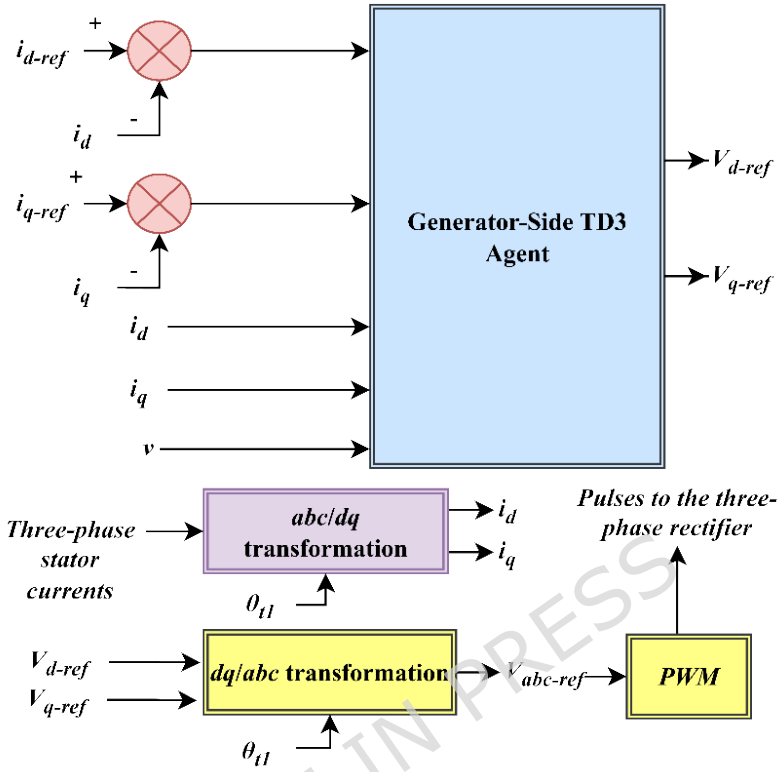


Fig. 3. Control loop of the generator side using the TD3 agent.

Also, a reward function (r_{gen}) based on the minimization of the *dq* currents errors is formulated using Eqs. (36) - (38) to guide the agent in the training:

$$r_{id} = 0.03 - |e_{id}| \tag{36}$$

$$r_{iq} = 0.03 - |e_{iq}| \tag{37}$$

$$r_{gen} = r_{id} + r_{iq} \tag{38}$$

Moreover, the most important training parameters are given in Table 2. These values are selected after many trial-and-error sessions. The agent is trained in MATLAB Simulink for 10 s. The results of the training procedure are illustrated in Fig. 4.

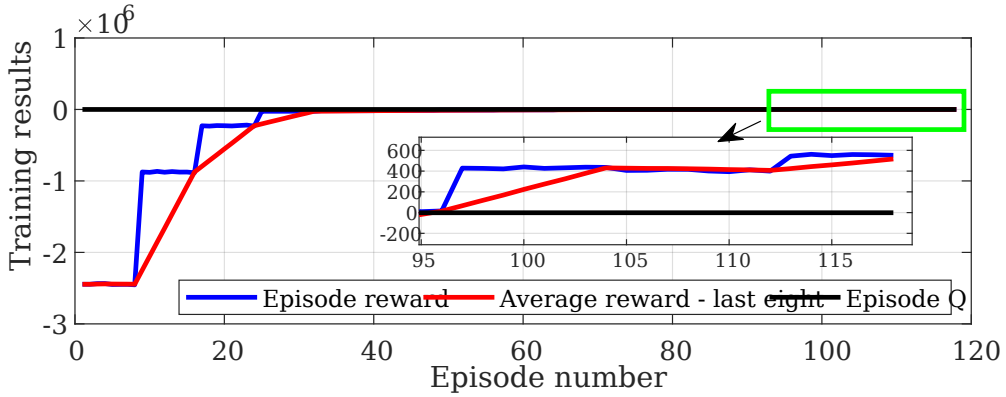


Fig. 4. Results of training the generator-side TD3 agent.

Table 2. The generator-side agent training options.

Parameter	Value
Actor network	2 hidden layers with 64 and 128 neurons (ReLU function) with an output layer (tanh function).
Critic network	2 hidden layers with 128 and 256 neurons (ReLU function).
Agent sample time	1×10^{-4}
α_a	1×10^{-5}
α_Q	1×10^{-3}
γ	0.99
Mini-batch size	1024
Experience buffer length	3×10^6
τ	0.005
Policy update frequency	2
Gaussian noise - action taken	Standard deviation = 0.5, and decay rate = 1×10^{-6}
Gaussian noise - target action	Standard deviation = 0.2
Parallelization	8 parallel workers - async mode.
Wall-clock training time	3 hours
Total environment steps	12.8×10^6 in 128 episodes

3.4. Control of the three-phase inverter using PI controllers

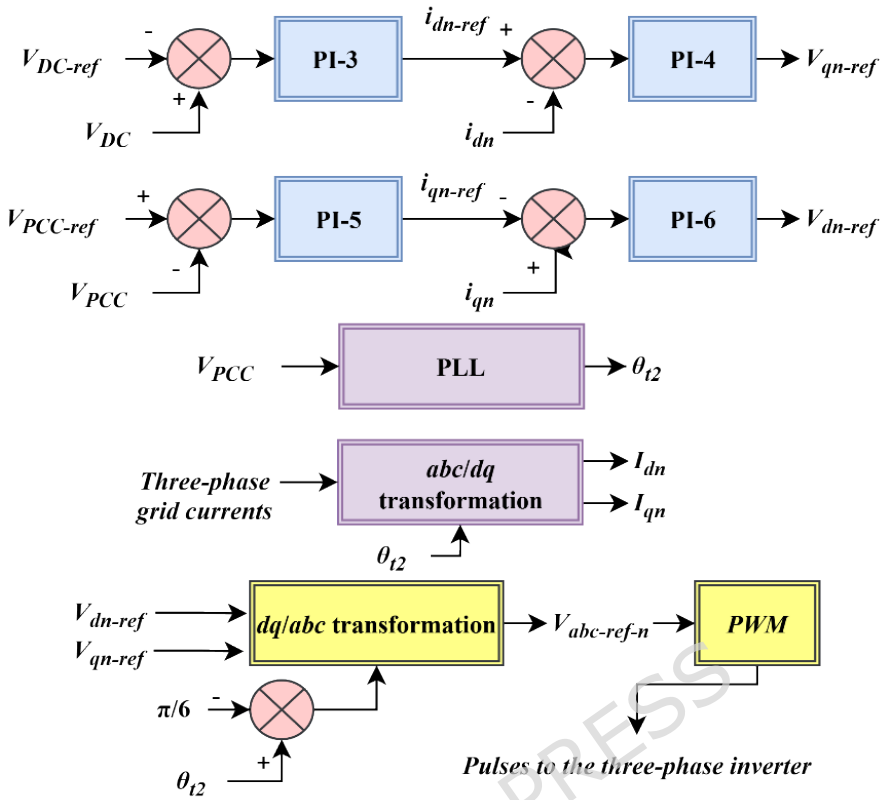


Fig. 5. Control loop of the two-level inverter.

The controller regulates V_{DC} and V_{PCC} at 1 pu. This is achieved by controlling the grid-side *dq* currents (i_{dn} & i_{qn}). These currents are responsible for adjusting the grid's active and reactive powers (P_{grid} and Q_{grid}), and hence, V_{DC} and V_{PCC} could be controlled. The classical control system is formed of four PI controllers. The first controller's input is the error in V_{DC} ($e_{V_{DC}}$) which in turn gives the reference direct axis current ($i_{dn\text{-}ref}$). The error between $i_{dn\text{-}ref}$ and i_{dn} is given to another controller that gives the reference quadrature axis voltage ($V_{qn\text{-}ref}$). Similarly, the error in V_{PCC} ($e_{V_{PCC}}$) is provided to another controller that generates the reference quadrature axis current ($i_{qn\text{-}ref}$). The error between $i_{qn\text{-}ref}$ and i_{qn} is an input to another controller that gives the reference direct axis voltage ($V_{dn\text{-}ref}$). $V_{dn\text{-}ref}$ and $V_{qn\text{-}ref}$ are converted into *abc* voltages ($V_{abc\text{-}ref\text{-}n}$) using the inverse Clarke-Park transformation. The transformation angle (θ_{t2}) is calculated using a phase-locked loop (PLL). 30° is subtracted from θ_{t2} due to different connections of the primary and the secondary windings of the transformer. Finally, the switching actions are provided to the IGBTs by comparing $V_{abc\text{-}ref\text{-}n}$ with a 1-kHz triangular waveform. A summary of the control

loop of the inverter is depicted in Fig. 5. Each PI controller has two gains (k_p and k_i). The gains are presented by $G = [k_{p1}\ k_{i1}\ k_{p2}\ k_{i2}\ ....]$, where the proportional and integral gains of the *i*-th PI controller are denoted by k_{pi} and k_{ii}, respectively. These gains were tuned in[38] and $G = [627.69\ 717.1\ 800\ 1000\ 7.71\ 80\ 1.2\ 22.5\ 3.66\ 219.3\ 2.2\ 25]$.

3.5. Control of the three-phase inverter using a TD3 agent

There are two TD3-agent approaches proposed for the grid-side controller. The first one is the replacement of the four PI controllers with a TD3 agent that takes several observations as an input and provides the reference grid-side *dq* voltages ($V_{dn\text{-}ref\text{-}1}$ and $V_{qn\text{-}ref\text{-}1}$). The selected inputs are $e_{V_{DC}}$, $e_{V_{PCC}}$, i_{dn}, i_{qn}, and θ_{t2}. The new control loop is illustrated in Fig. 6.

The reward function ($r_{grid\text{-}voltages}$) selected for training this agent is given in Eqs. (39) - (41):

$$r_{V_{DC}} = 0.005 - |e_{V_{DC}}| \tag{39}$$

$$r_{V_{PCC}} = 0.005 - |e_{V_{PCC}}| \tag{40}$$

$$r_{grid\text{-}voltages} = 5(r_{V_{DC}} + r_{V_{PCC}}) \tag{41}$$

The agent is trained for 10 s under steady state. The training parameters are given in Table 3, and the training results are shown in Fig. 7.

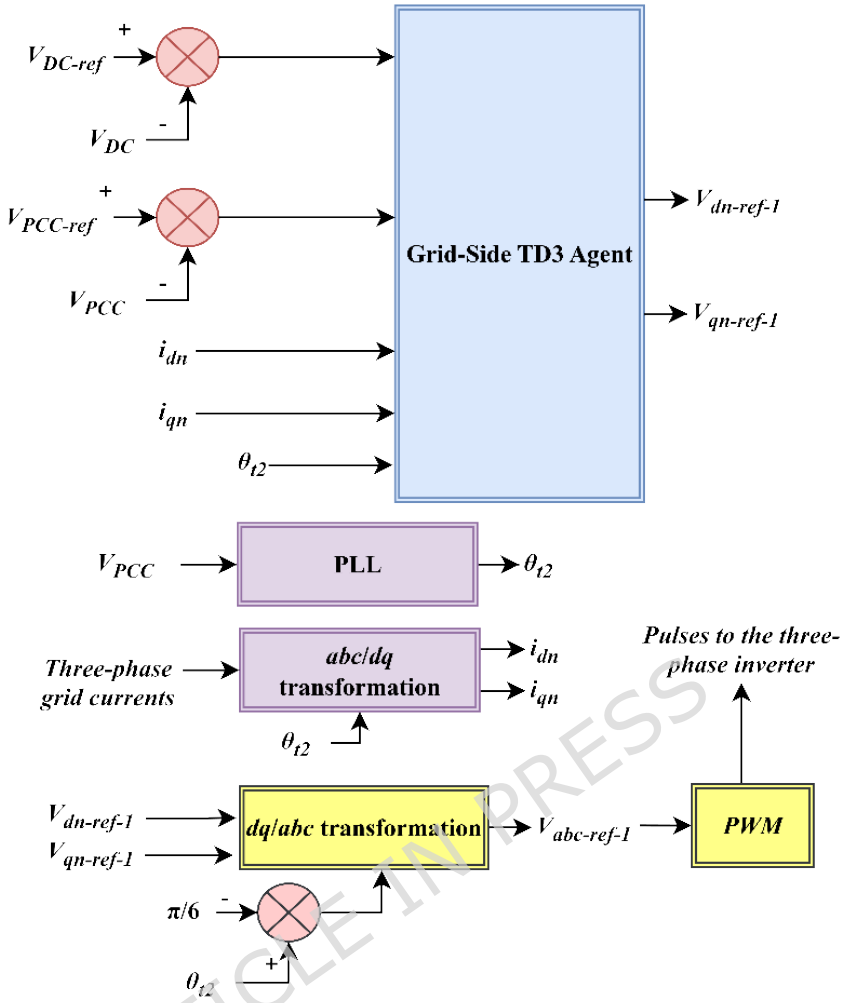


Fig. 6. New control loop of the grid side using the TD3 agent.

Table 3. The grid-side agent training options.

Parameter	Value
Actor network	2 hidden layers with 32 and 64 neurons (ReLU) with an output layer (tanh).
Critic network	2 hidden layers with 64 and 64 neurons (ReLU).
Agent sample time	4×10^{-4}
α_a	5×10^{-6}
α_Q	1×10^{-5}
γ	0.995
Experience buffer length	2×10^{6}
Mini-batch size	1024
τ	0.005
Policy update frequency	2
Gaussian noise - action taken	Standard deviation = 0.2 and decay rate = 1×10^{-5}
Gaussian noise - target action	Standard deviation = 0.2

Parallelization	8 parallel workers - async mode.
Wall-clock training time	16.5 hours
Total environment steps	18.8×10^6 in 752 episodes

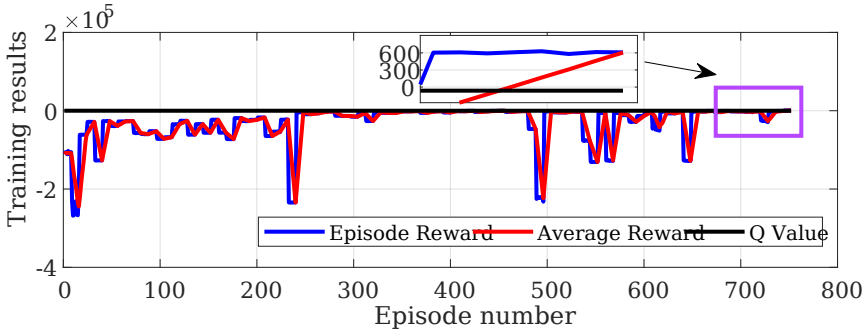


Fig. 7. Results of training the grid-side TD3 agent.

In the second approach, a hybrid PI-TD3 agent is deployed. In this control strategy, the first two PI controllers that provide the reference grid *dq* currents are kept as they are, and the second pair of PI controllers is exchanged with a TD3 agent. This control loop is given in Fig. 8. The agent's control objective is to minimize the errors in *dq* currents. Therefore, the reward function ($r_{\text{grid-currents}}$) for training this agent is proposed in Eqs. (42) - (44):

$$r_{i_{dn}} = 0.01 - |e_{i_{dn}}| \tag{42}$$

$$r_{i_{qn}} = 0.01 - |e_{i_{qn}}| \tag{43}$$

$$r_{\text{grid-currents}} = 5(r_{i_{dn}} + r_{i_{qn}}) \tag{44}$$

The agent is trained for 10 s under the 3LG condition to help the actor acquire a strong policy that suppresses the *dq* currents during fault. The training parameters are given in Table 4, and the training results are shown in Fig. 9.

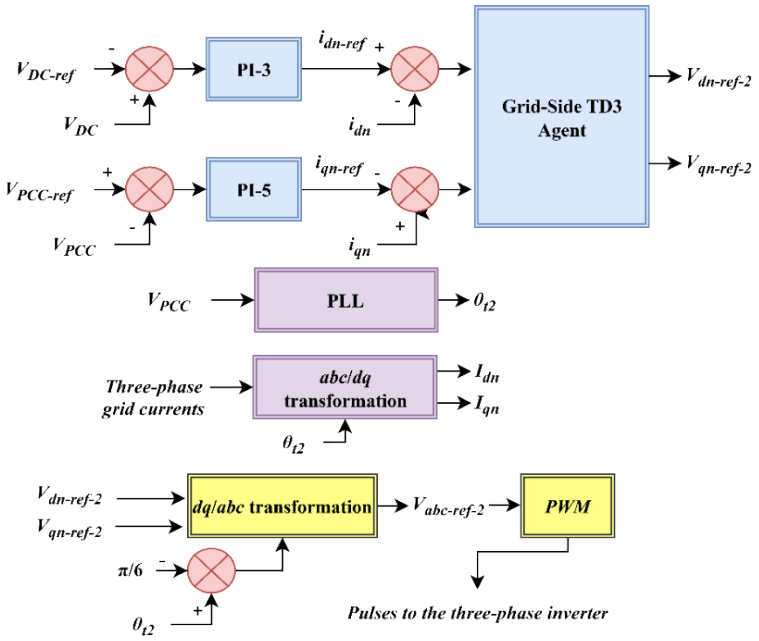


Fig. 8. The hybrid PI-TD3 agent control loop.

Table 4. The hybrid PI-TD3 agent training options.

Parameter	Value
Actor network	2 hidden layers with 18 and 32 neurons (ReLU) with an output layer (tanh).
Critic network	2 hidden layers with 64 and 128 neurons (ReLU).
Agent sample time	1×10^{-4}
α_a	1×10^{-5}
α_Q	1×10^{-4}
γ	0.995
Experience buffer length	2×10^{6}
Mini-batch size	1024
τ	0.005
Policy update frequency	2
Parallelization	8 parallel workers - async mode.
Gaussian noise - action taken	Standard deviation = 0.2 and decay rate = 1×10^{-6}
Gaussian noise - target action	Standard deviation = 0.2
Wall-clock training time	1.5 hours
Total environment steps	6.4×10^{6} in 64 episodes

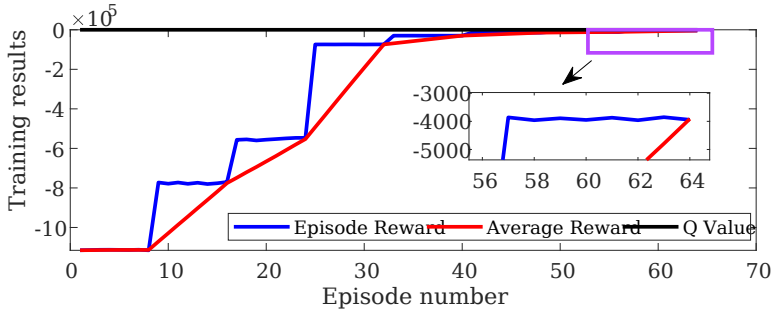


Fig. 9. Results of training the grid-side hybrid PI-TD3 agent.

The summary of these agents, including observations, actions, and the corresponding reward functions, is provided in Table 5.

Table 5. Summary of the agents

Agent	Observations vector	Actions vector	Reward function
Generator-side TD3	$[e_{id}, e_{iq}\ i_d, i_q, v]$	$[V_{d\text{-}ref}, V_{q\text{-}ref}]$	$r_{gen} = r_{id} + r_{iq}$
PI-TD3	$[e_{V_{DC}}, e_{V_{PCC}}, i_{dn}, i_{qn}, \theta_{t2}]$	$[V_{dn\text{-}ref\text{-}1}, V_{qn\text{-}ref\text{-}1}]$	$r_{grid\text{-}voltages} = 5(r_{V_{DC}} + r_{V_{PCC}})$
Hybrid PI-TD3	$[e_{i_{dn}}, e_{i_{qn}}]$	$[V_{dn\text{-}ref\text{-}2}, V_{qn\text{-}ref\text{-}2}]$	$r_{grid\text{-}currents} = 5(r_{i_{dn}} + r_{i_{qn}})$

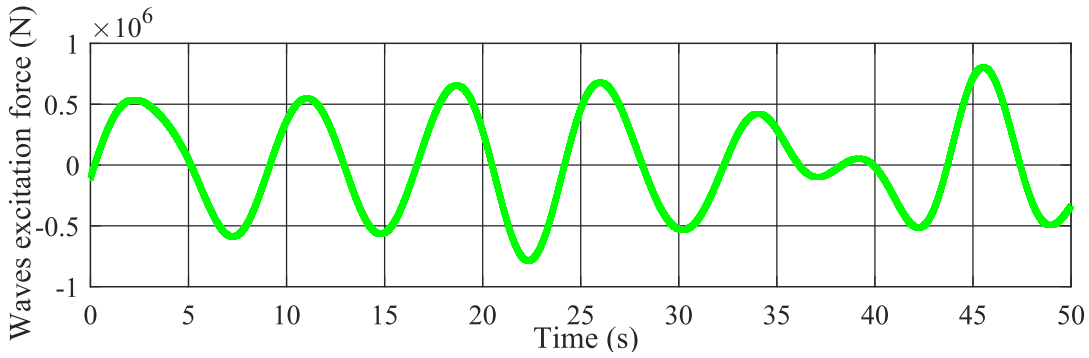
In the next section, these controllers are investigated under steady state (untrained states) and transient state (fault conditions).

4. TD3 agents' results under various operating conditions

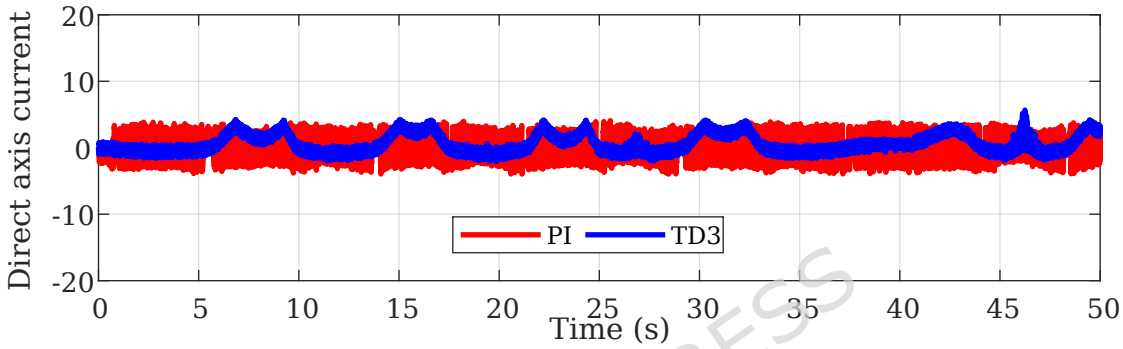
In this part, the double-agent configuration is analyzed during steady and transient states under various grid faults. The system is executed in MATLAB/Simulink 2024b with a fixed step solver type ode4 and a solver time step of 10 µs. The converters use non-ideal IGBT/diode switches with an on-state resistance of 1 mΩ.

4.1. Generator-side agent performance

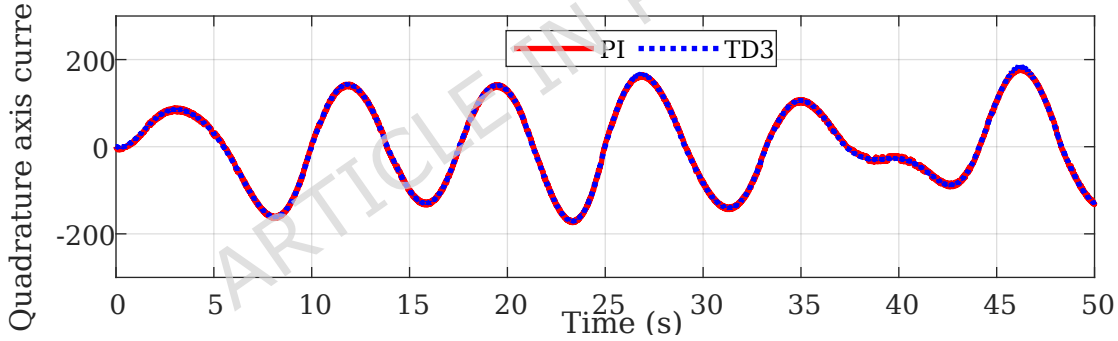
The TD3 agent of the generator side is trained for 10 s in MATLAB Simulink. The system runs for an additional 40 s, involving states the agent hasn't encountered during training. The results of the system's simulation under the effect of irregular sea waves (H_s = 4 m and T_p = 8 s) are depicted in Fig. 10(a)-(f). These include F_e, i_d, i_q, v and x, P_{gen}, and E_{gen}. The performance of the PI controllers and the TD3 agent is benchmarked against each other.



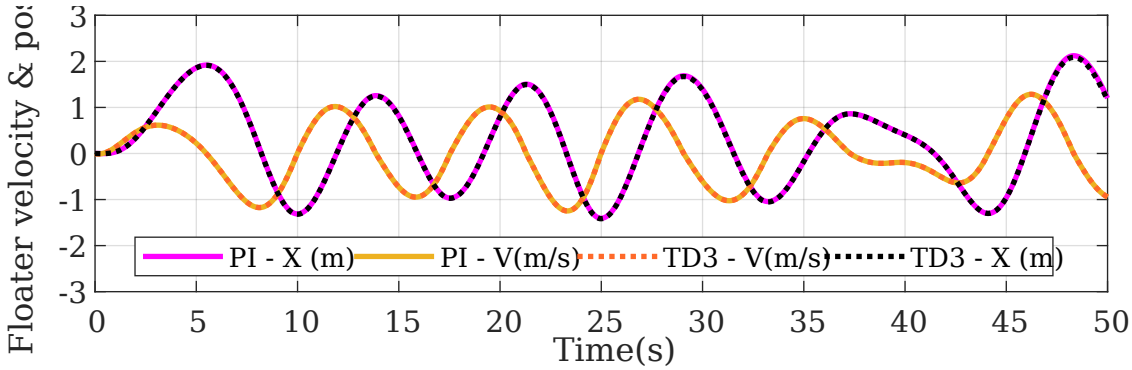
(a)



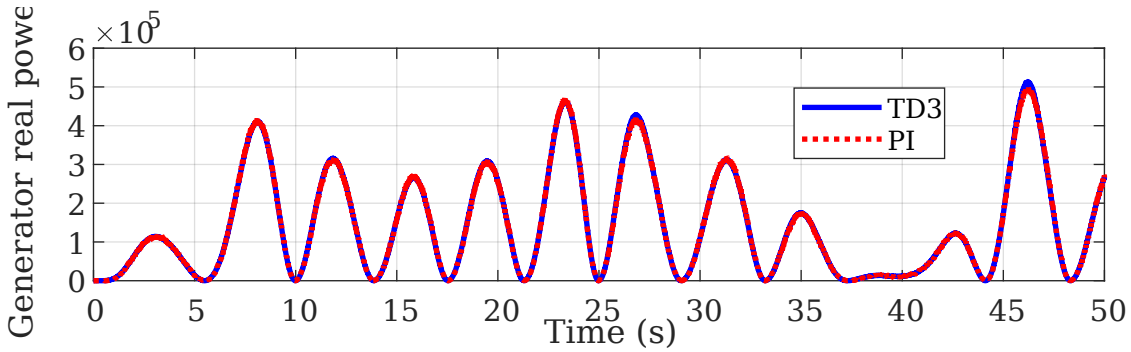
(b)



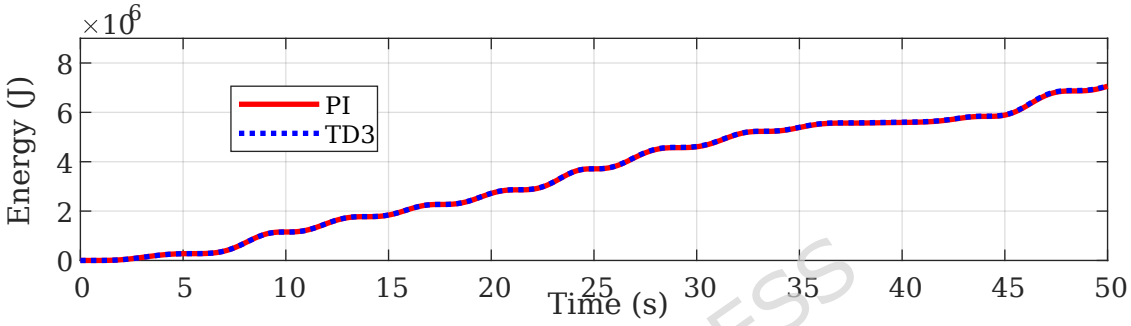
(c)



(d)



(e)



(f)

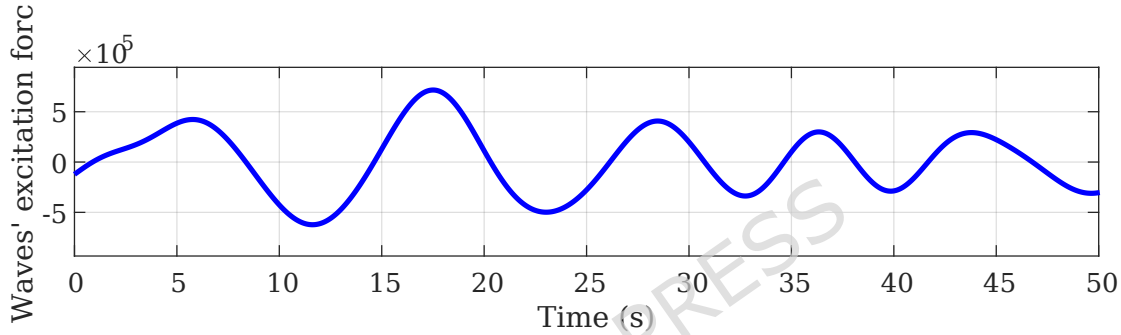
Fig. 10. The PI controllers and the TD3 agent performance under irregular waves characterized by H_s = 4 m and T_p = 8 s, (a) F_e, (b) i_d, (c) i_q, (d) v and x, (e) P_{gen}, and (f) E_{gen}

The results emphasize that the TD3 agent generator-side controller and the PI controllers achieve the control objectives of reducing the currents' errors. i_d is fluctuating around zero and i_q is tracked by the TD3 agent to maximize the energy production from sea waves. Additionally, the following table provides the quantitative comparison between the TD3 agent's performance and the PI controllers:

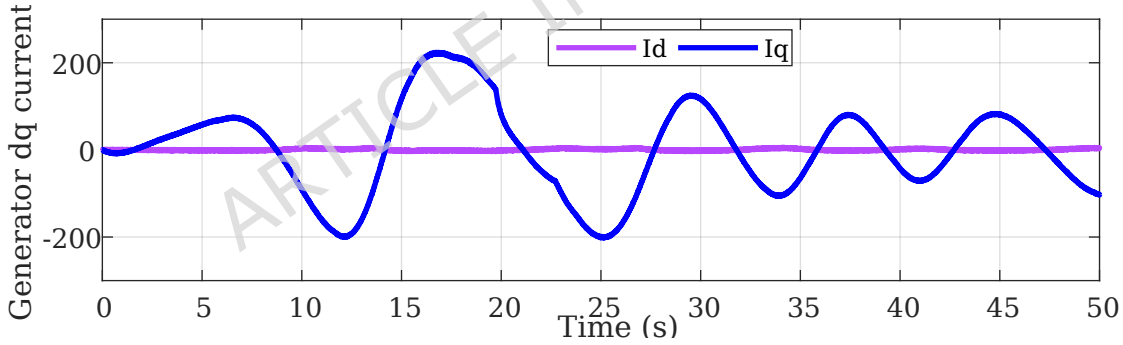
Table 6. Generator-side TD3 comparison with classical PI controllers.

Metric	PI	TD3 agent	Best
Integral squared error (ISE) of i_d	0.5796	0.434	TD3 agent
ISE of i_q	1.584	0.426	TD3 agent
Average stator losses	3.912 kW	3.946 kW	PI controllers
Average generated power	141.2 kW	141.54 kW	TD3 agent
Energy produced by the generator (E_{gen})	7.058 MJ	7.077 MJ	TD3 agent

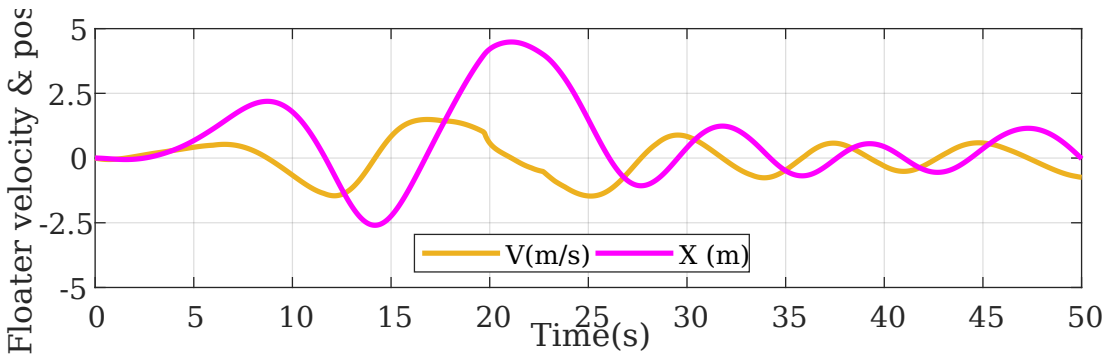
Table 6 confirms the higher control efficiency of the TD3 agent compared to the PI controllers. The ISE for the *dq* currents is lower, and the average generated power and energy over 50 seconds are higher with the TD3 agent. It should be noted that the TD3 agent has stator losses that are 0.034 kW higher than those of the PI controllers. However, the TD3 agent's average generated power is 0.34 kW higher. To further confirm the TD3 agent's tracking reliability, the control efficacy was investigated across multiple sea wave states and with a 20% increase in the floater's mass (to simulate biofouling). The results are given in Figs. 11-13. Additionally, Table 7 provides the ISE for the *dq* currents under these scenarios.



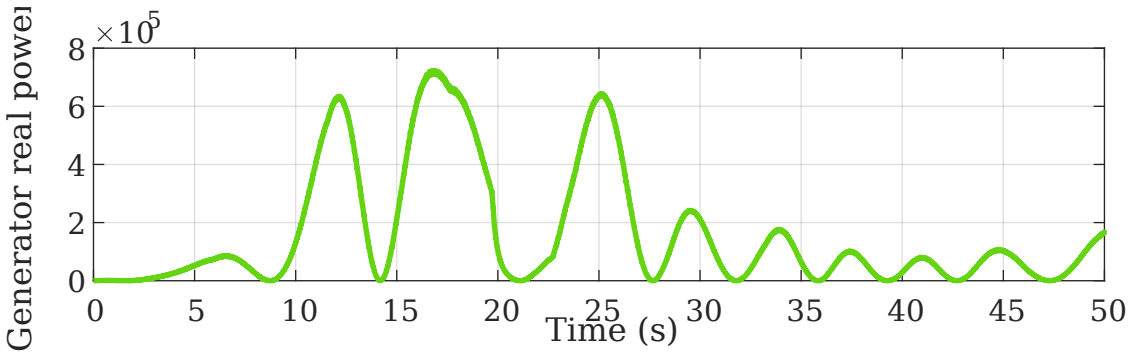
(a)



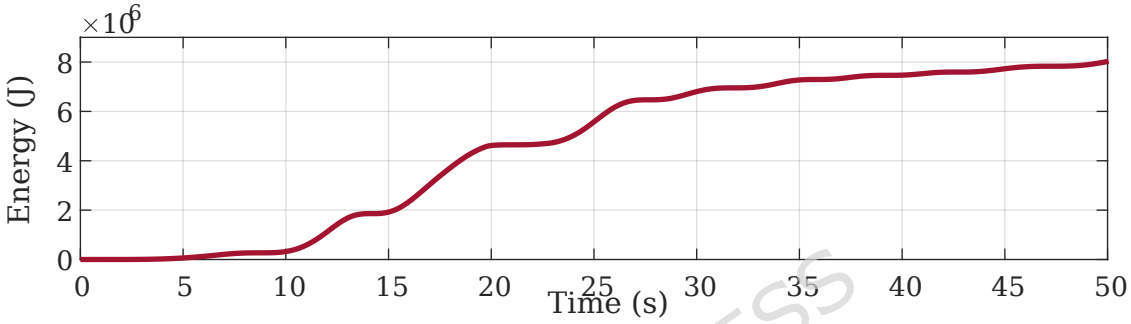
(b)



(c)

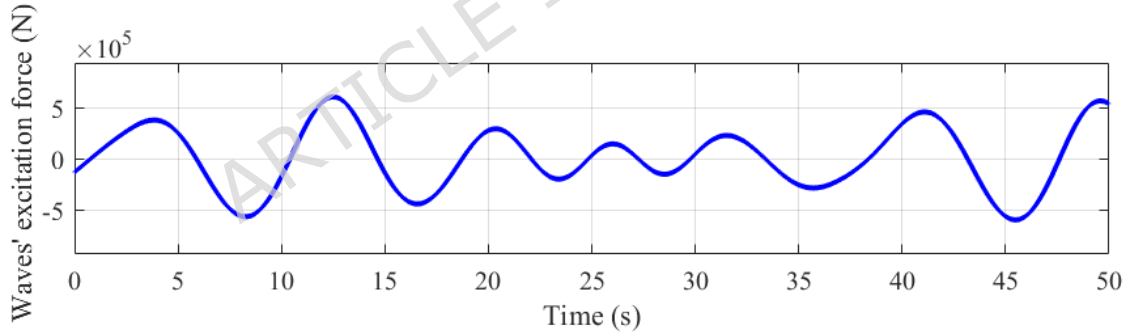


(d)

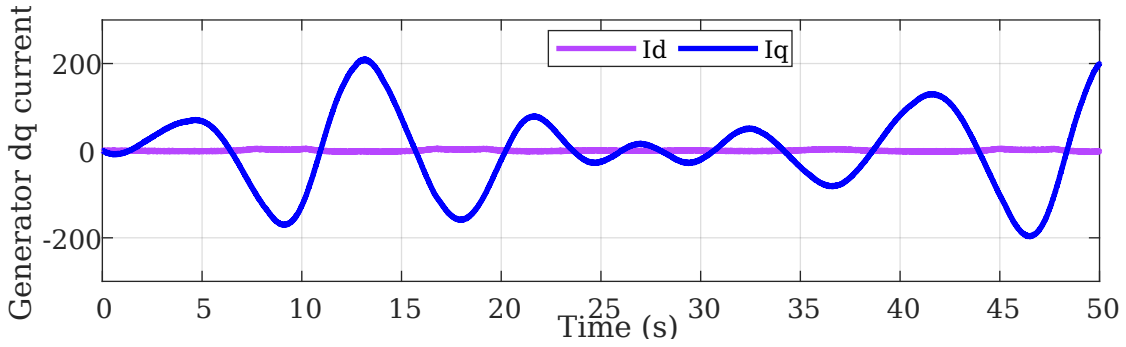


(e)

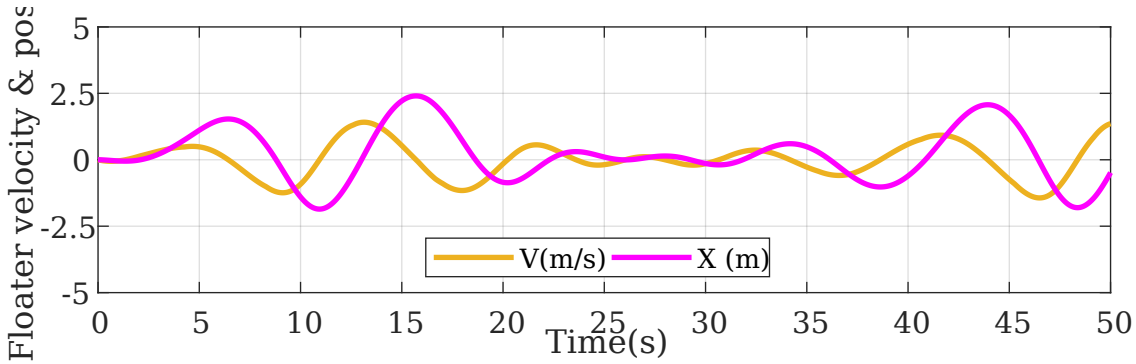
Fig. 11. The TD3 agent performance under irregular waves characterized by H_s = 2.1 m and T_p = 12 s, (a) F_e, (b) i_d and i_q, (c) v and x, (d) P_{gen}, and (e) E_{gen}



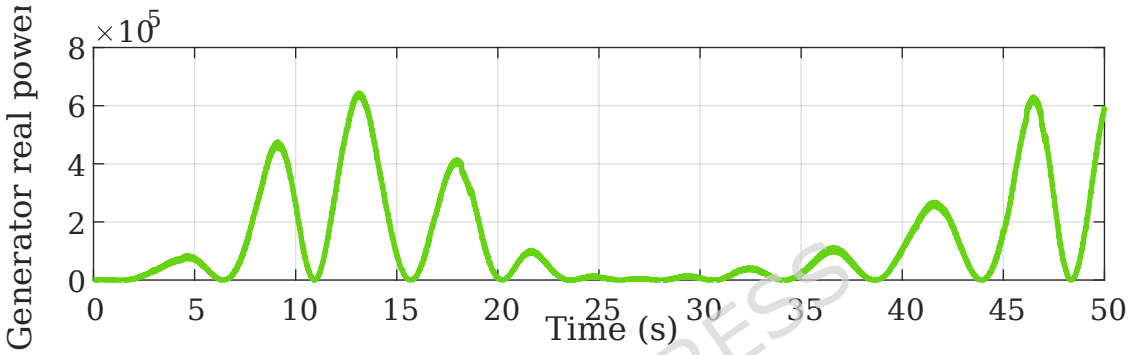
(a)



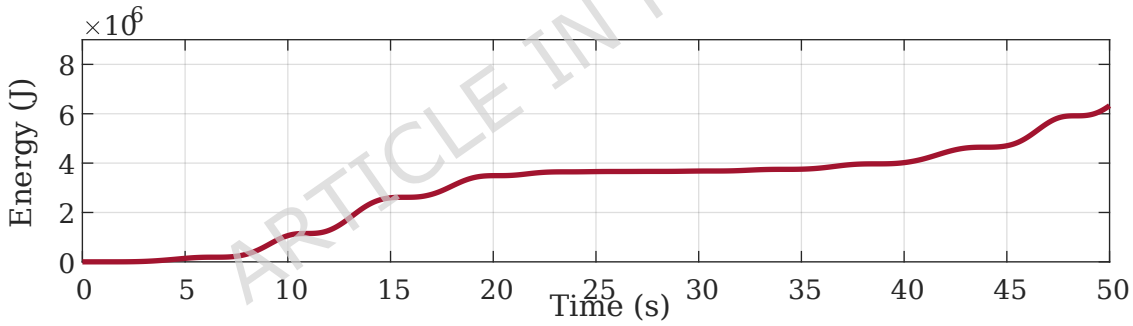
(b)



(c)

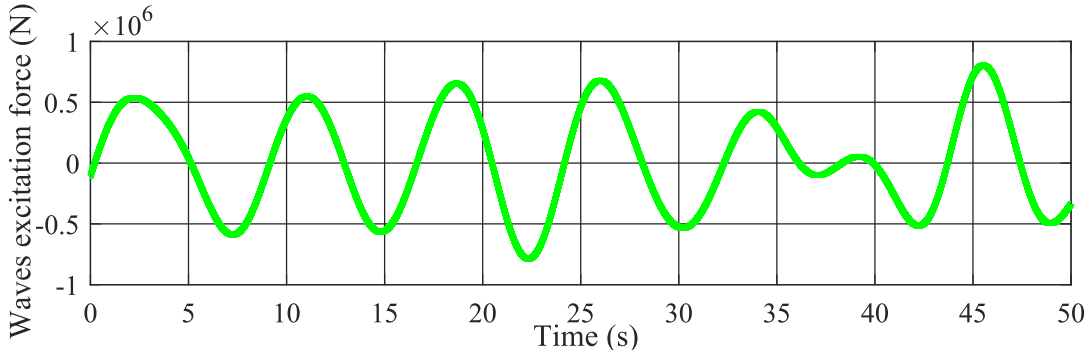


(d)

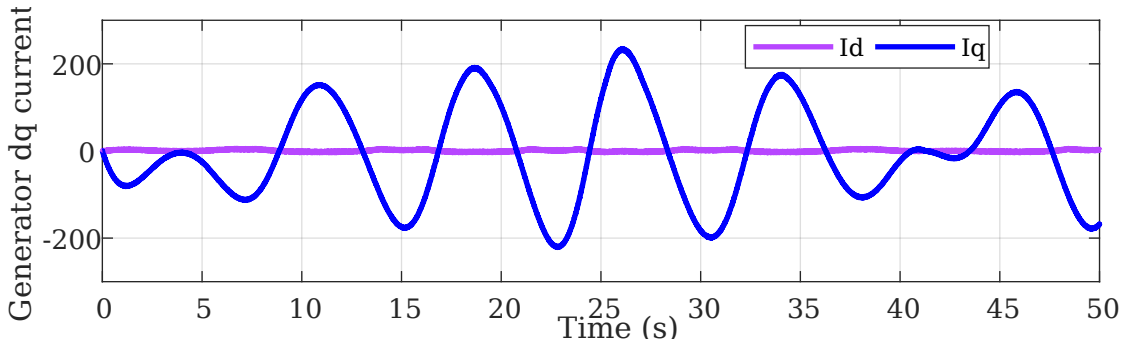


(e)

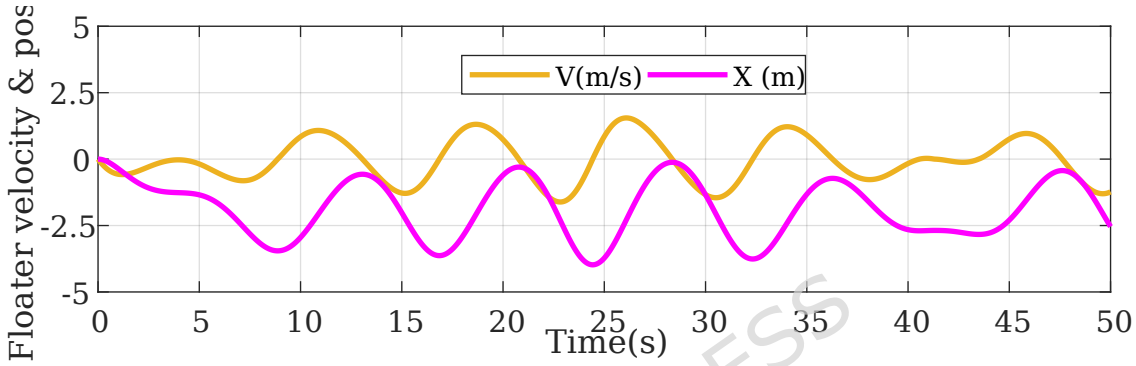
Fig. 12. The TD3 agent's performance under irregular waves characterized by H_s = 2.53 m and T_p = 8.58 s, (a) F_e, (b) i_d and i_q, (c) v and x, (d) P_{gen}, and (e) E_{gen}



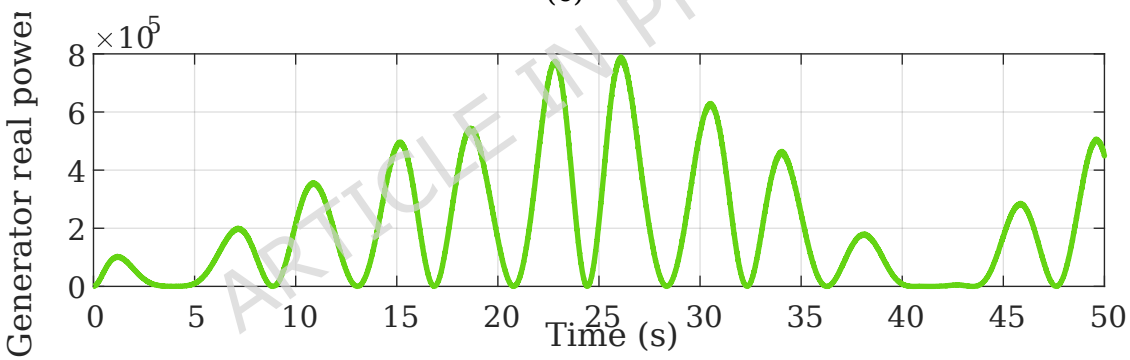
(a)



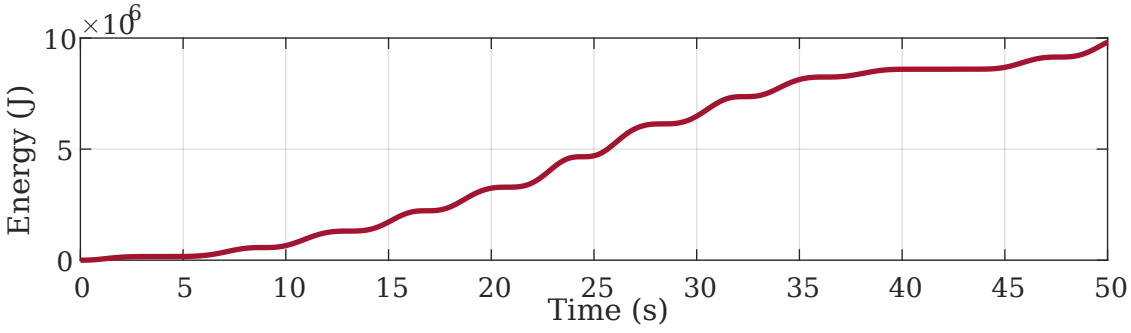
(b)



(c)



(d)



(e)

Fig. 13. The TD3 agent performance under irregular waves characterized by H_s = 4 m and T_p = 8 s and 20% increase in the floater mass (a) F_e, (b) i_d and i_q, (c) v and x, (d) P_{gen}, and (e) E_{gen}

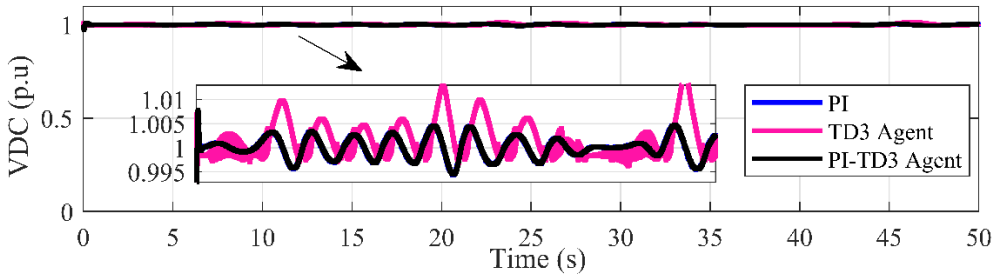
Table 7. Generator-side TD3 agent - *dq* currents ISE under different conditions.

Metric	i_d - ISE	i_q - ISE
Irregular waves characterized by H_s = 2.1 m and T_p = 12 s	0.746	2.64
Irregular waves characterized by H_s = 2.53 m and T_p = 8.58 s	0.69	1.06
20% increase in the floater mass	0.975	2.055

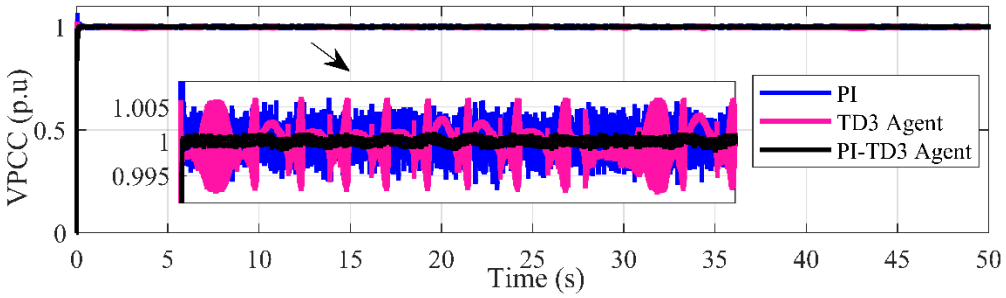
Table 7 shows that the TD3 agent maintained the ISE of the currents within reasonable values across different sea states, including the scenario of increasing the floater mass due to the attachment of marine organisms to the AWS floater.

4.2. Grid-side agent performance under steady state

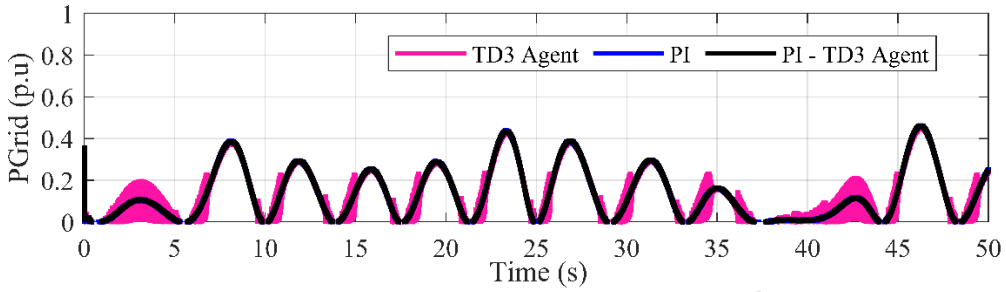
The TD3 agent and the hybrid-PI agent are trained for 10 s. The simulation runs for another 40 s to evaluate their performance under untrained states. It is benchmarked against the classical PI configuration. The simulation results include V_{DC}, V_{PCC}, grid active and reactive powers (P_{grid} and Q_{grid}), and grid-side *dq* currents, as shown in Fig. 14(a)-(f). Furthermore, Table 8 provides a comparison between V_{PCC}, V_{DC}, I_{dn}, and I_{qn} results obtained in the steady state condition. These are the primary objectives of the various control loops.



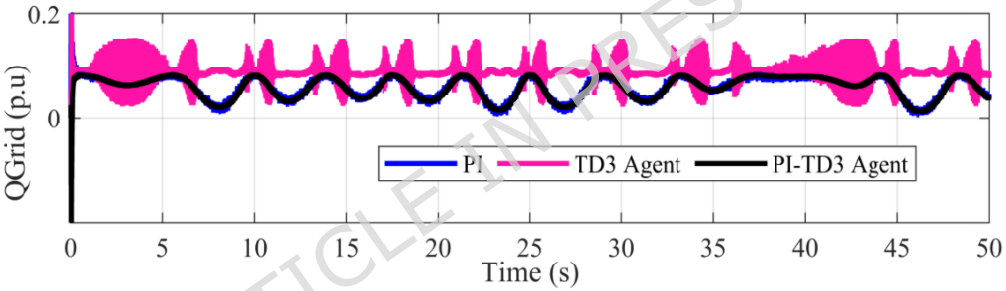
(a)



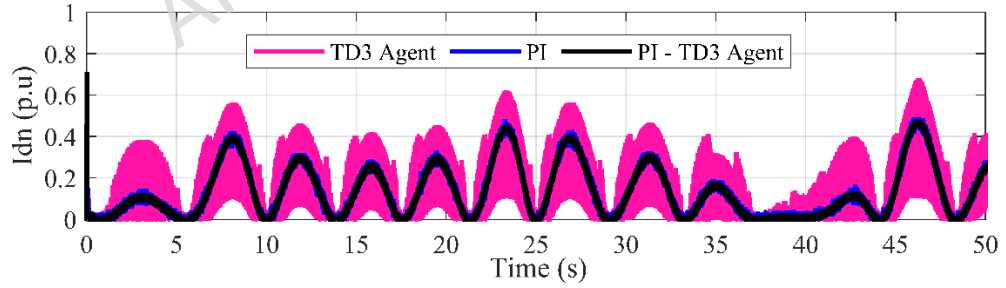
(b)



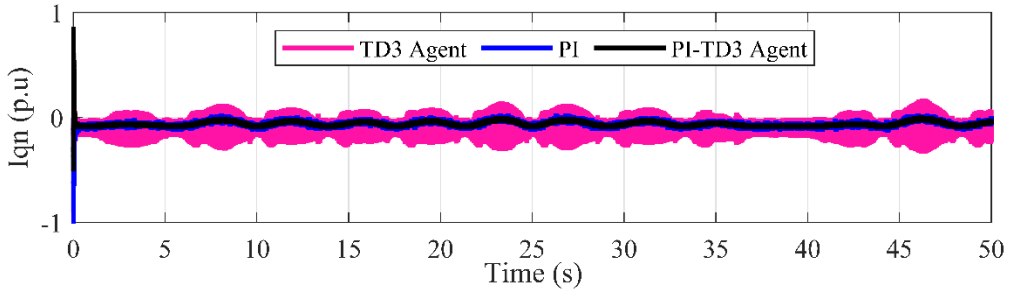
(c)



(d)



(e)



(f)

Fig. 14. (a) V_{DC}, (b) V_{PCC}, (c) P_{grid}, (d) Q_{grid}, (e) i_{dn}, and (f) i_{qn}.

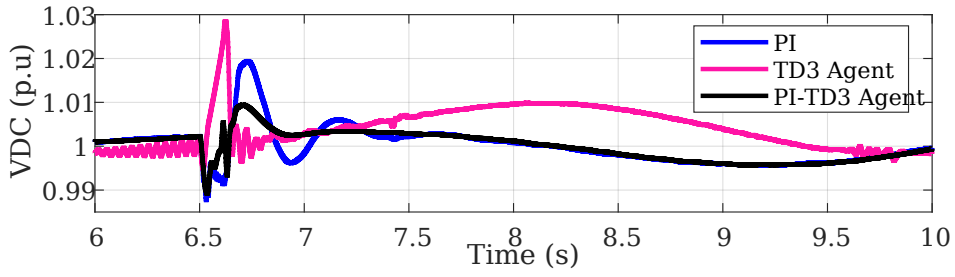
Table 8. Steady state results comparison.

Metric	PI	TD3 agent	Hybrid PI-TD3	Best
V_{PCC} – ISE	4.4382	3.524	3.73	TD3 agent followed by the hybrid PI-TD3
V_{DC} – ISE	0.256	0.865	0.2654	PI controllers and hybrid PI-TD3
V_{PCC} – peak-to-peak ripple	0.012 pu	0.012 pu	10^{-3} pu	Hybrid PI-TD3
V_{DC} – peak-to-peak ripple	0.0105 pu	0.017 pu	0.01 pu	Hybrid PI-TD3 and PI controllers
V_{PCC} – mean value	0.9998	0.9995	0.9998	TD3 agent
V_{DC} – mean value	1.0001	1.0016	1.0001	Hybrid PI-TD3 and PI controllers
V_{PCC} – settling time (5%)	0.042 seconds	0.02 seconds	0.06 seconds	TD3 agent
V_{PCC} – standard deviation value	0.0094	0.0084	0.0086	TD3 agent
V_{DC} – standard deviation value	0.0023	0.0038	0.0023	Hybrid PI-TD3 and PI controllers
I_{dn} – ISE	23.355	Not applicable	5.6802	Hybrid PI-TD3
I_{qn} – ISE	23.58	Not applicable	8.1842	Hybrid PI-TD3

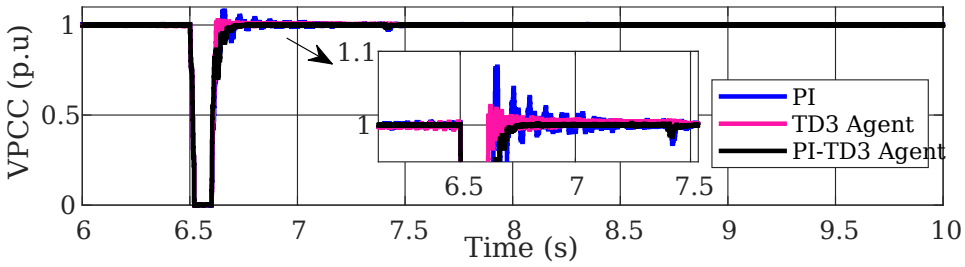
The steady-state results emphasize that the three controllers were able to stabilize V_{DC} and V_{PCC} around 1 pu. However, in Fig. 14(b), the hybrid PI-TD3 agent has the lowest fluctuations for V_{PCC} compared with others. Furthermore, the single TD3-agent approach is considered aggressive when controlling these voltages, resulting in waveform fluctuations. This requires a strong filter to remove those switching harmonics. However, this aggressive behavior leads to the minimum settling time for V_{PCC}. Additionally, the hybrid-PI TD3 offers the best-balanced performance in terms of ISE of V_{PCC} and V_{DC} + the lowest peak-peak ripple for V_{PCC}.

4.3. Grid-side agent performance under transient state

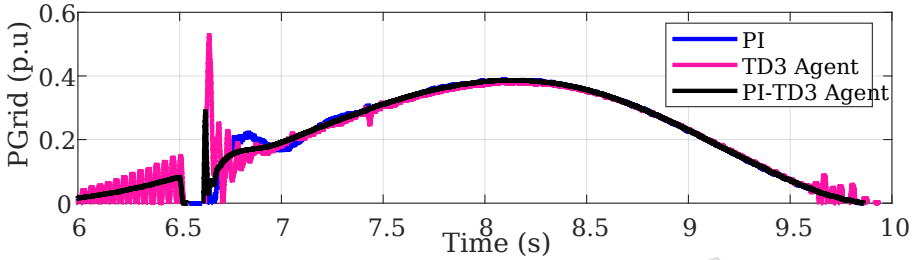
The system is exposed to different grid faults like LLLG (three-line to ground fault), 2LG (two phases A and B + ground fault), LL (line to line fault between phase A and phase B), LG (phase A with ground fault), and LLL (three-line to line fault) faults to assess the performance of the PI controllers, TD3 agent, and the hybrid PI–TD3 agent during these transient states. These faults were examined at point "F" in Fig. 1. Each fault occurs at one of the parallel transmission lines at $t = 6.5$ s. The faults are cleared by the circuit breakers associated with the transmission line at $t = 6.6$ s. The breakers successfully reclosed at $t = 7.4$ s. The fault resistance, grounding resistance, and breaker resistance are 0.1 Ω, 0.01 Ω, and 0.01 Ω, respectively. The breaker logic complies with the LVRT requirements. The results indicate that breaker trips within 100 ms because V_{PCC} drops below 0.5 pu in 99% of scenarios during the various grid faults. The system responses include V_{DC}, V_{PCC}, P_{grid} and Q_{grid}, and grid-side *dq* currents in Figs. 15–19. Finally, a quantitative analysis between V_{DC}, V_{PCC}, P_{grid}, Q_{grid} , and the grid-side *dq* currents are specified in Table 9.



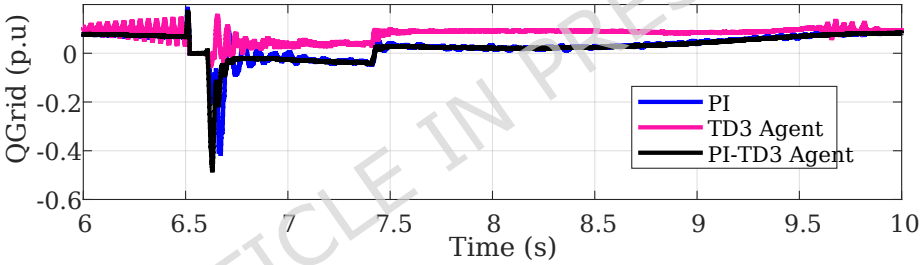
(a)



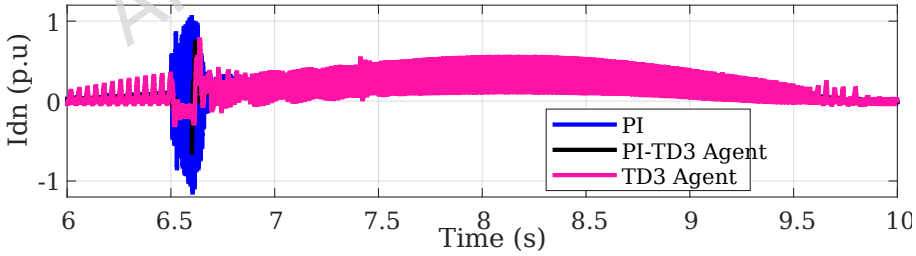
(b)



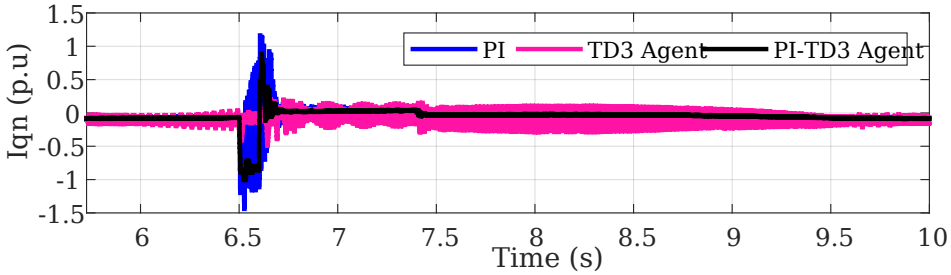
(c)



(d)

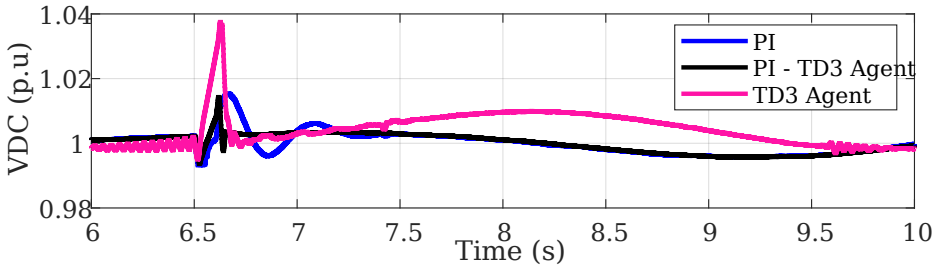


(e)

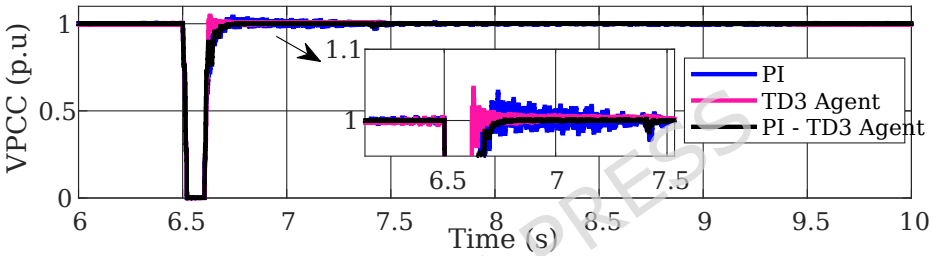


(f)

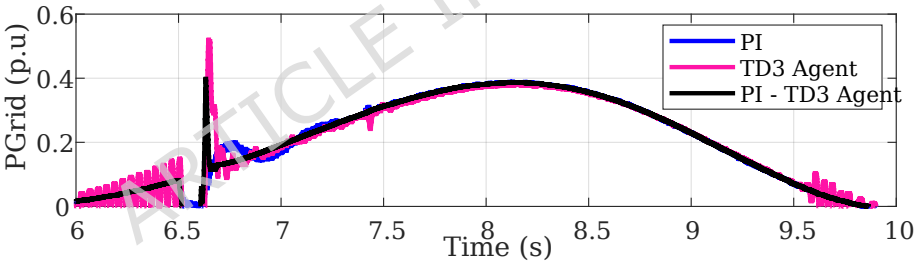
Fig. 15. System dynamic response under a symmetrical LLLG fault. The subplots (a–f) show that the proposed TD3 controllers maintained the DC-link voltage, PCC voltage, and grid currents within stable limits during the fault.



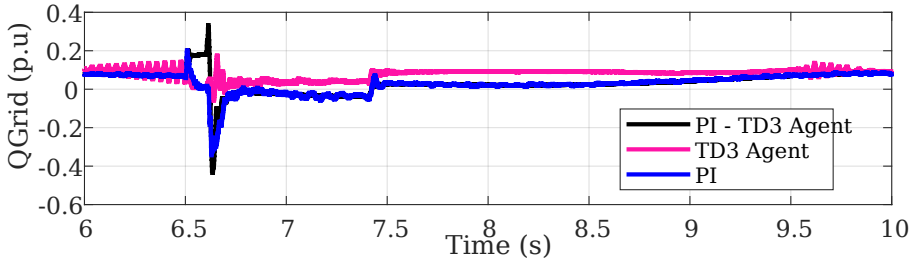
(a)



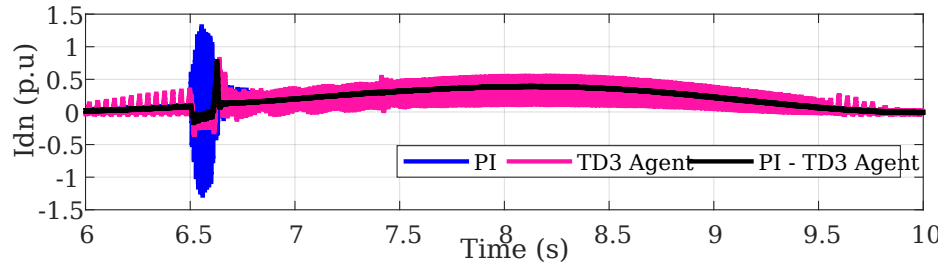
(b)



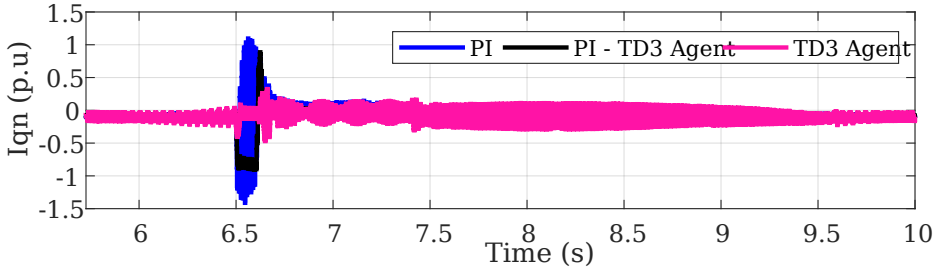
(c)



(d)

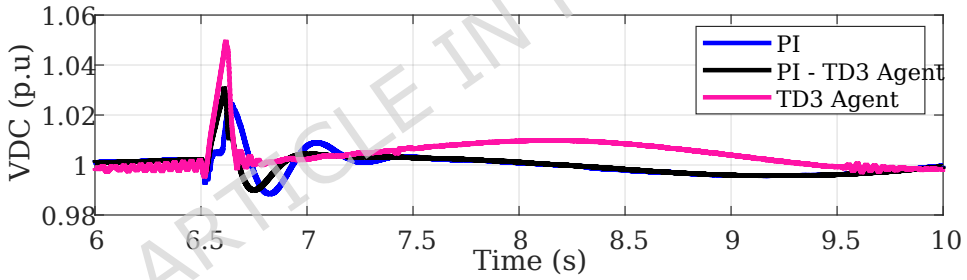


(e)

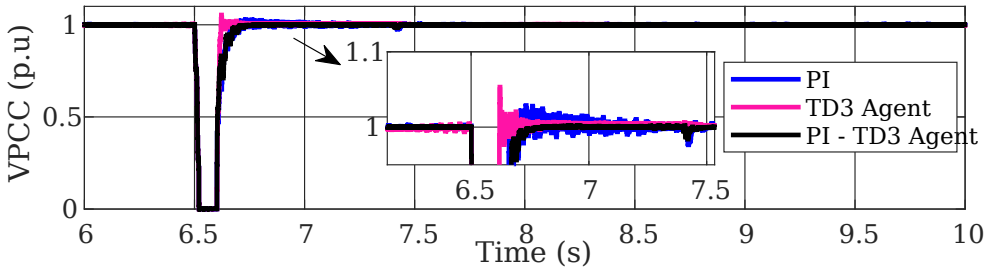


(f)

Fig. 16. Performance comparison during an asymmetrical 2LG fault. The results illustrate the robustness of the TD3 agent in handling unbalanced conditions compared to the PI controllers.



(a)



(b)

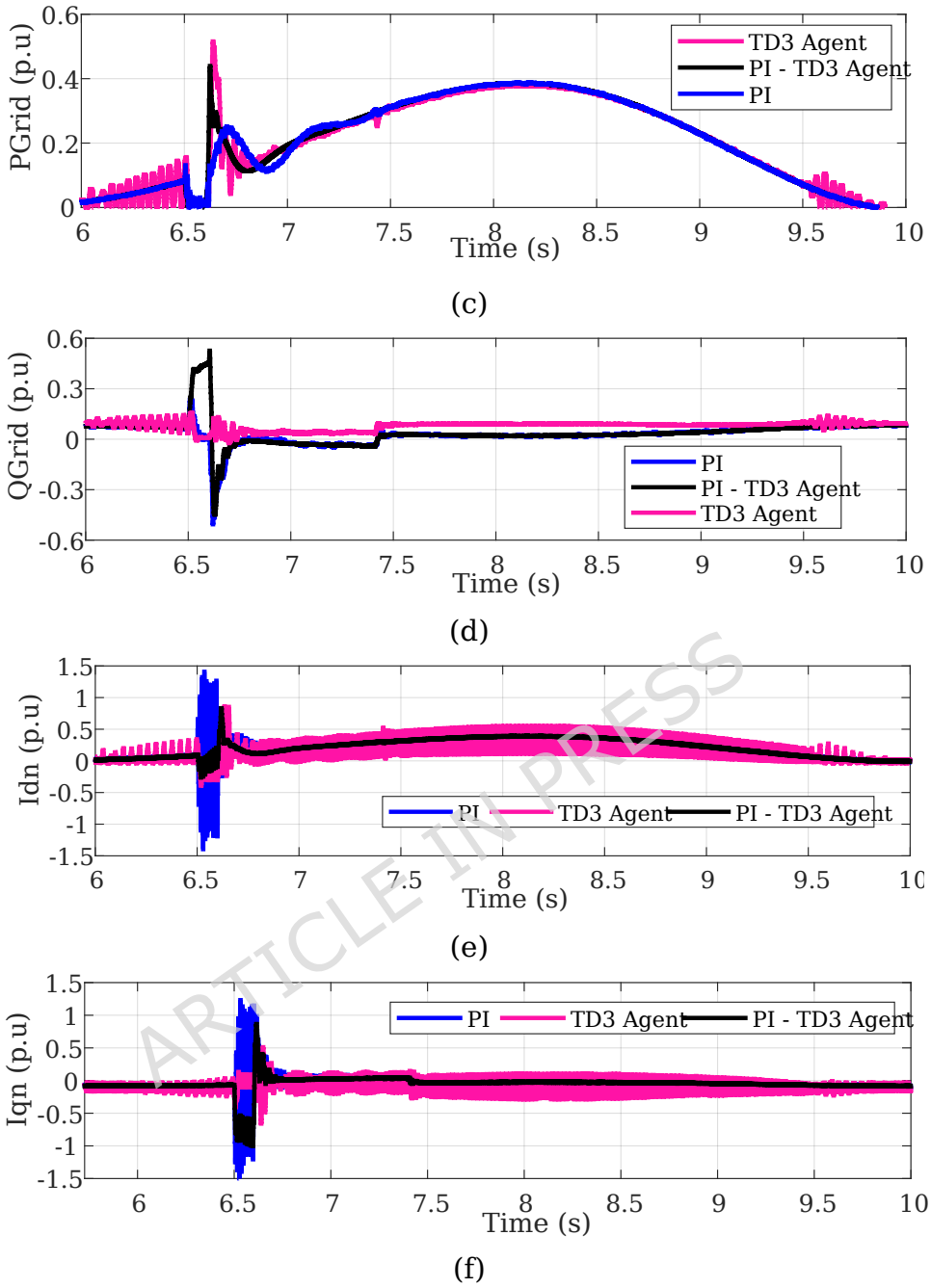
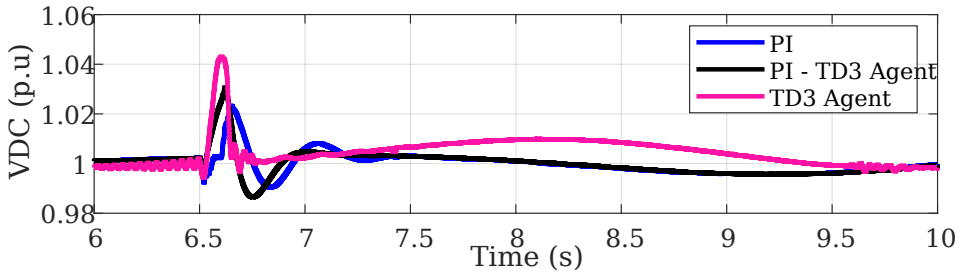
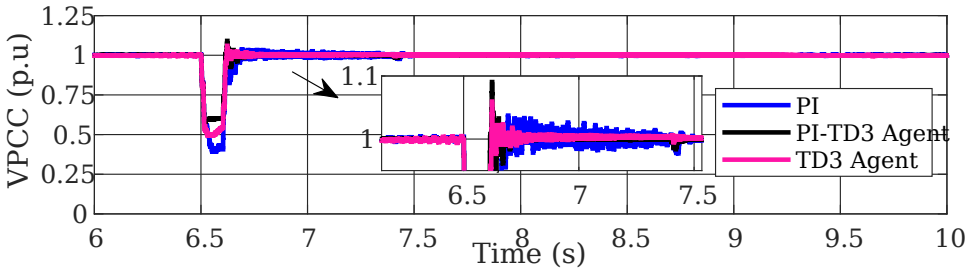


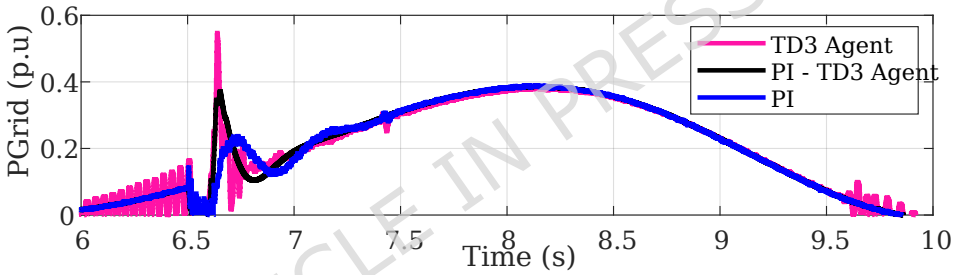
Fig. 17. Control effectiveness under an LL fault scenario. The subplots confirm that the *dq* grid currents didn't reach saturation limits during the disturbance in the case of the TD3 agents.



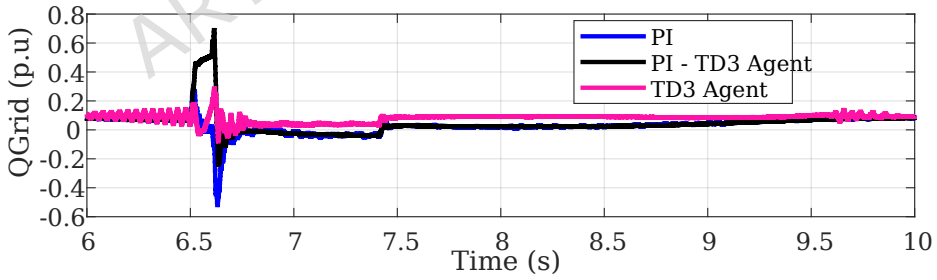
(a)



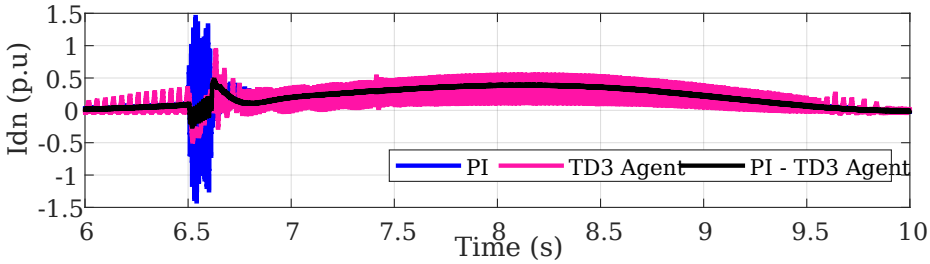
(b)



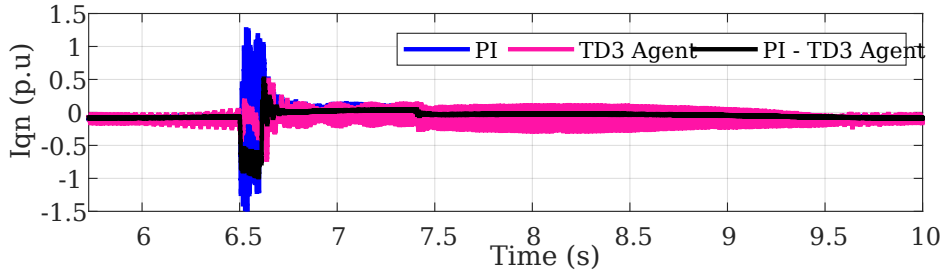
(c)



(d)

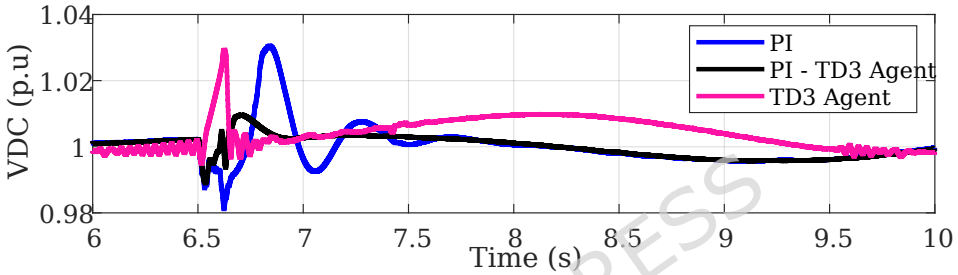


(e)

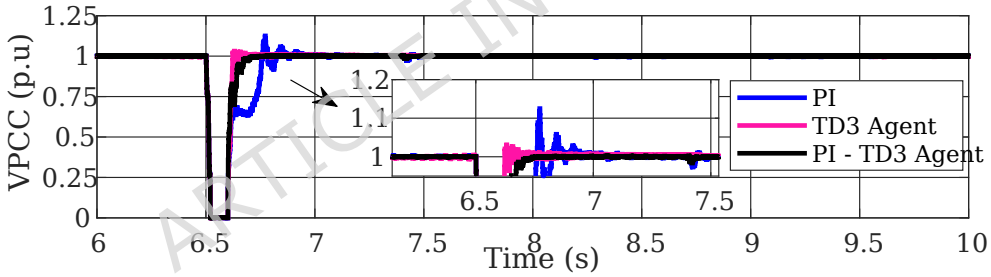


(f)

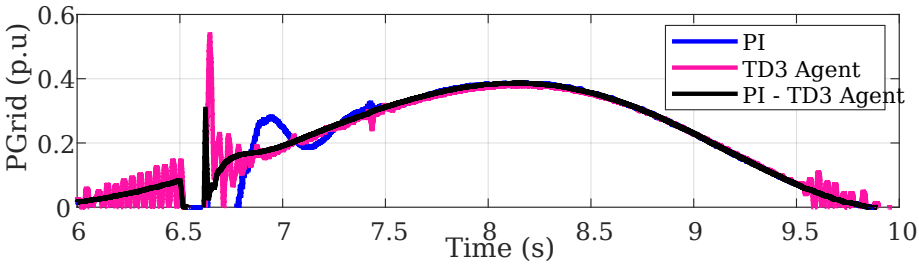
Fig. 18. Transient behavior under an LG fault at the PCC. The rapid recovery of all system variables (a–f) confirms the effectiveness of the TD3 control policy.



(a)



(b)



(c)

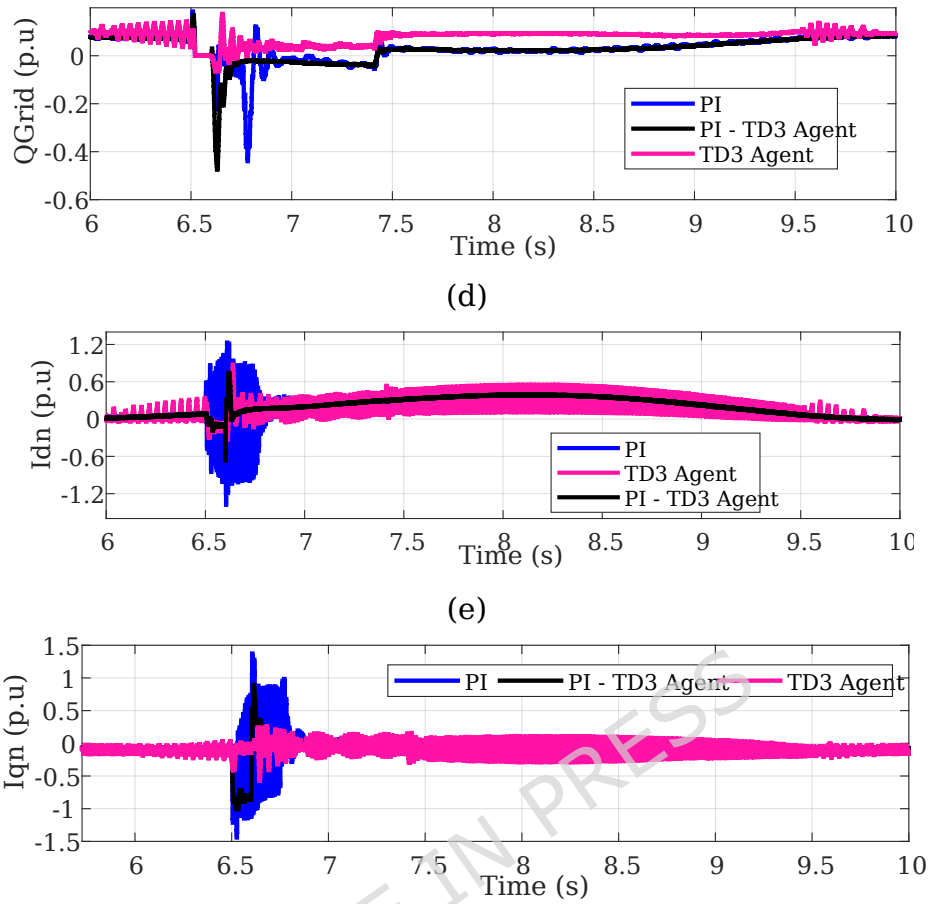


Fig. 19. Impact of an LLL fault on system stability. The TD3 agents ensured that active/reactive power, *dq* currents, V_{DC}, and V_{PCC} returned to steady state quickly.

Table 9. Comparison between the controllers' results.

Metric	PI	TD3 agent	Hybrid PI-TD3	Comment
V_{PCC} – maximum overshoot and minimum voltage values	LLLG: 1.08 & 0. 2LG: 1.04 & 0. LL: 1.03 & 0. LG: 1.04 & 0.4. LLL: 1.12 & 0.	LLLG: 1.025 & 0. 2LG: 1.046 & 0. LL: 1.053 & 0. LG: 1.06 & 0.5. LLL: 1.03 & 0.	LLLG: 1.001 & 0. 2LG: 1.002 & 0. LL: 1.001 & 0. LG: 1.09 & 0.56. LLL: 1.003 & 0.	Hybrid PI-TD3 has the lowest maximum overshoot in most of the scenarios.
V_{DC} – maximum overshoot and undershot	LLLG: 1.02 & 0.988. 2LG: 1.015 & 0.993. LL: 1.024 &	LLLG: 1.028 & 0.996. 2LG: 1.037 & 0.995. LL: 1.049 &	LLLG: 1.009 & 0.99. 2LG: 1.014 & 0.994.	TD3 agent has the lowest undershoot. Hybrid PI-TD3 has the lowest overshoot.

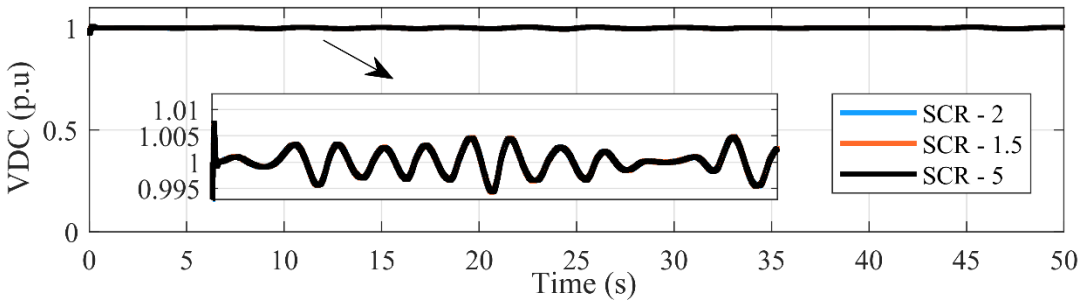
voltage values	0.988. LG: 1.023 & 0.99. LLL: 1.03 & 0.981. LLLG: 0.22. 2LG: 0.2. LL: 0.25. LG: 0.237. LLL: 0.28.	0.995. LG: 1.043 & 0.995. LLL: 1.029 & 0.995. LLLG: 0.53. 2LG: 0.52. LL: 0.52. LG: 0.55. LLL: 0.54.	LL: 1.03 & 0.99. LG: 1.03 & 0.987. LLL: 1.01 & 0.989. LLLG: 0.29. 2LG: 0.4. LL: 0.44. LG: 0.37. LLL: 0.3.	
P_{grid} - peak				PI controller has the lowest peak of P_{grid} during fault.
Q_{grid} – max. and min. values	LLLG: 0.183 & -0.41. 2LG: 0.03 & -0.34. LL: 0.233 & -0.5. LG: 0.27 & -0.52. LLL: 0.184 & -0.44.	LLLG: 0.11 & -0.05. 2LG: 0.18 & -0.06. LL: 0.16 & -0.01. LG: 0.29 & -0.07. LLL: 0.17 & -0.06.	LLLG: 0.17 & -0.48. 2LG: 0.33 & -0.43. LL: 0.53 & -0.45. LG: 0.69 & -0.24. LLL: 0.17 & -0.48.	Hybrid PI-TD3 and provides the highest reactive power during fault. However, the TD3 agent has the lowest absorption of reactive power from the grid.
Integral absolute error (IAE) for V_{PCC} & V_{DC}	LLLG: 158.91. 2LG: 155.04. LL: 151.22. LG: 108.16. LLL: 178.88.	LLLG: 151.24. 2LG: 156.16. LL: 150.97. LG: 104.67. LLL: 151.82.	LLLG: 133.45. 2LG: 136.75. LL: 135.87. LG: 75.55. LLL: 133.63.	Hybrid PI-TD3 has the lowest IAE for the control objectives.
Absolute peak values for I_{dn} & I_{qn} (pu) and the peak true RMS current of phase A (pu)	LLLG: 1.14, 1.16, 1.22. 2LG: 1.311, 1.41, 2.54. LL: 1.4, 1.23, 1.39. LG: 1.44, 1.5, 3.49. LLL: 1.37, 1.44, 1.28.	LLLG: 0.771, 0.5, 0.489. 2LG: 0.81, 0.46, 3.15. LL: 0.86, 0.66, 0.83. LG: 0.93, 0.711, 2.42. LLL: 0.876, 0.575, 0.528.	LLLG: 0.714, 1, 0.99. 2LG: 0.77, 0.89, 3.477. LL: 0.832, 0.996, 0.828. LG: 0.47, 0.969, 2.78. LLL: 0.74, 1, 0.994.	Hybrid PI-TD3 and TD3 agent maintain the *dq* currents in the [-1 1] range. Unlike the PI controllers that exceed 1 pu by a large margin during multiple scenarios. The single TD3 agent has the minimum RMS value of phase A in all

				faults except in 2LG fault.
Settling time (2%) - V_{PCC}	LLLG: 0.273 seconds. 2LG: 0.213 seconds. LL: 0.08 seconds. LG: 0.4 seconds. LLL: 0.263 seconds.	LLLG: 0.058 seconds. 2LG: 0.05 seconds. LL: 0.024 seconds. LG: 0.072 seconds. LLL: 0.058 seconds.	LLLG: 0.099 seconds. 2LG: 0.084 seconds. LL: 0.1 seconds. LG: 0.1 seconds. LLL: 0.1 seconds.	TD3 agent has the lowest settling time followed by the hybrid PI-TD3 agent.
Settling time (2%) - V_{DC}	LLLG: - 2LG: - LL: 0.069 seconds. LG: 0.074 seconds. LLL: 0.3 seconds.	LLLG: 0.038 seconds. 2LG: 0.046 seconds. LL: 0.042 seconds. LG: 0.037 seconds. LLL: 0.039 seconds.	LLLG: - 2LG: - LL: 0.022 seconds. LG: 0.044 seconds. LLL: -	The hybrid PI-TD3 maintained the voltage within 2% even during fault conditions in most cases, followed by the PI controllers. Despite, V_{DC} exceeded 2% several times in the case of the single TD3 agent, the agent was able to quickly bring it back to the 2% range.
Post-fault recovery time - V_{PCC}	LLLG: 0.042 seconds. 2LG: 0.056 seconds. LL: 0.052 seconds. LG: 0.015 seconds. LLL: 0.152 seconds.	LLLG: 0.02 seconds. 2LG: 0.023 seconds. LL: 0.015 seconds. LG: 0.014 seconds. LLL: 0.021 seconds.	LLLG: 0.042 seconds. 2LG: 0.05 seconds. LL: 0.042 seconds. LG: 0.013 seconds. LLL: 0.042 seconds.	TD3 agent has the quickest recovery of V_{PCC} compared to others.

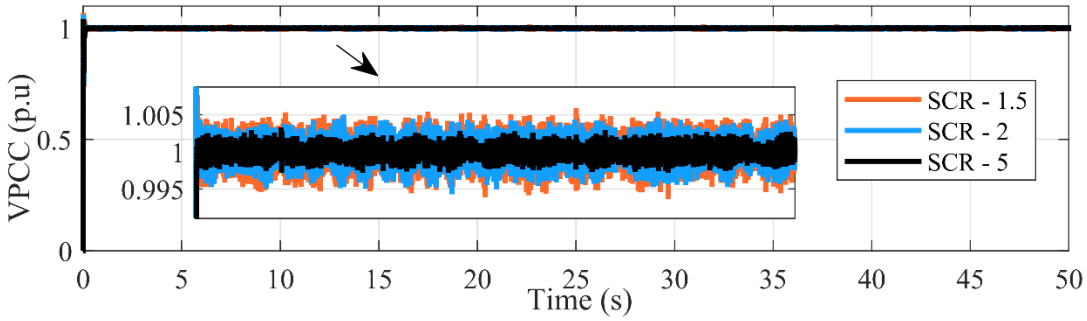
Under different fault conditions, grid-side *dq* currents are successfully suppressed under 1 pu by both the TD3 agent and the hybrid PI-TD3 agent. In contrast, the classical PI controllers failed to keep them below 1 pu, and in many cases, the currents reached 1.5 pu. The hybrid PI-TD3 and TD3 agents are considered the most effective two controllers compared to the classical PI controllers. The hybrid PI-TD3 has almost no overshoot in V_{PCC} and the lowest overshoot in V_{DC} during fault conditions. Second, the TD3 agent has the lowest undershoot in V_{DC}. While the PI control system has the lowest overshoot for the provided grid-active power under fault conditions. However, we are more concerned with providing reactive power. That's why the hybrid PI-TD3 outperformed the other controllers, as it has the highest provided Q_{grid} in most cases of grid faults. In summary, the suggested approach is a hybrid one that combines the TD3 agent with PI controllers to achieve optimal performance.

4.4. Testing the hybrid PI-TD3 grid-side agent under different grid short circuit ratios (SCR)

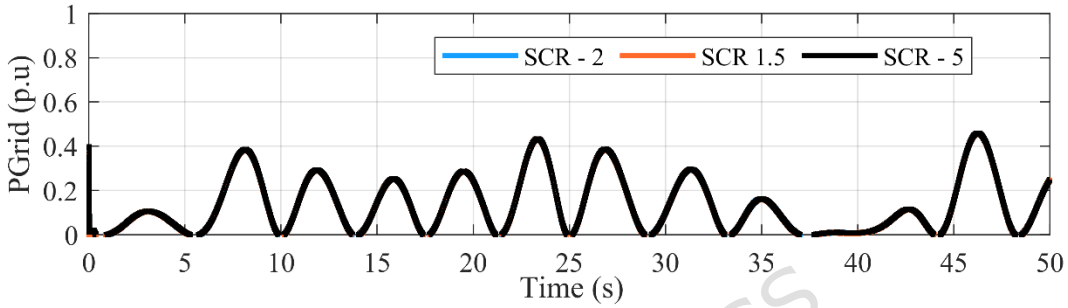
In this part, the performance of the hybrid PI-TD3 agent was investigated under different grid SCR ratios (1.5, 2, and 5). The SCR is the ratio between the short circuit MVA of the grid with respect to the rated generator power (1 MVA). This evaluates the behavior of the grid-connected system under weak and moderate power grids. The waveforms of the systems under steady-state are given in Fig. 20. Moreover, the system was also assessed under the effect of an LLLG fault in Fig. 21.



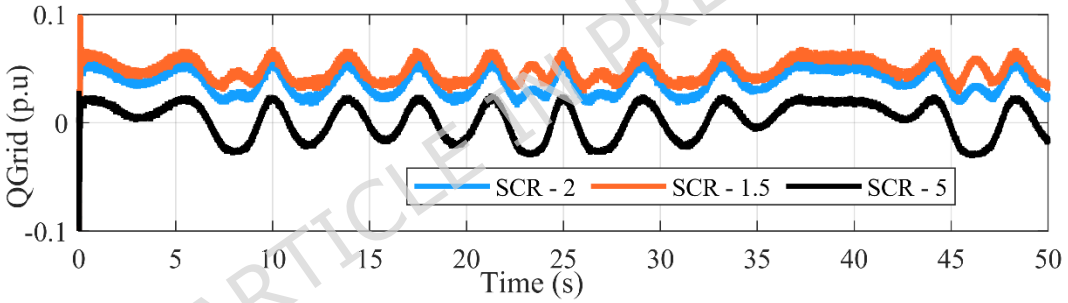
(a)



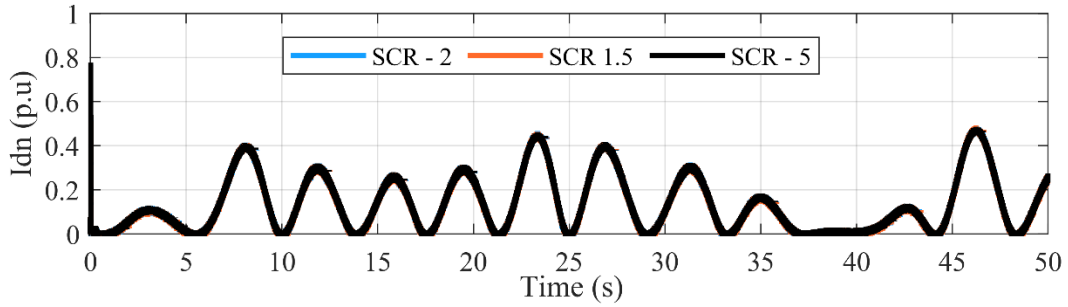
(b)



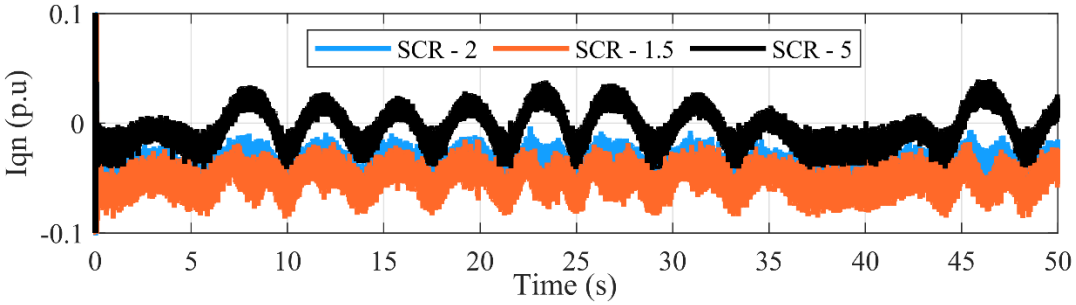
(c)



(d)

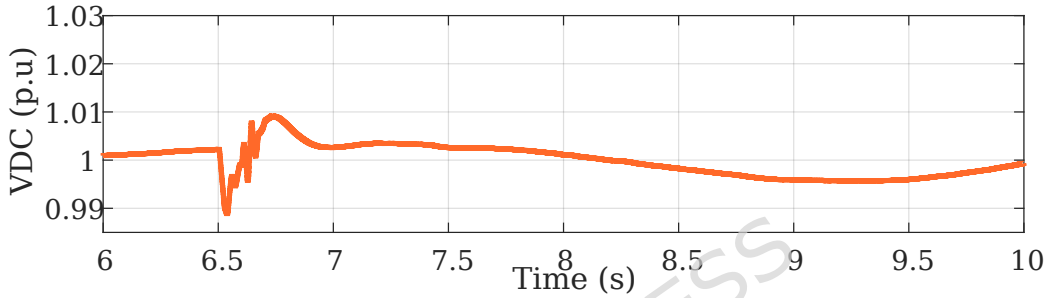


(e)

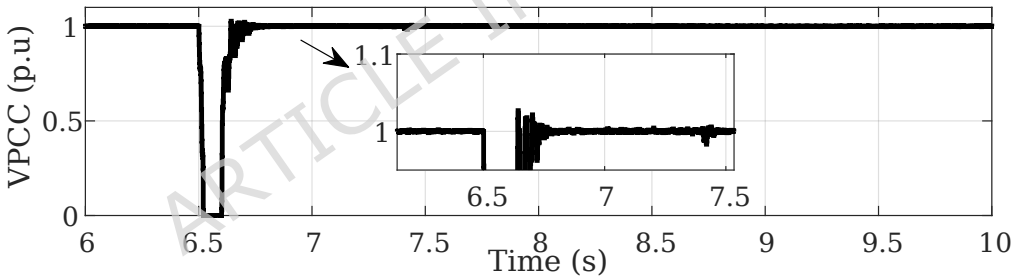


(f)

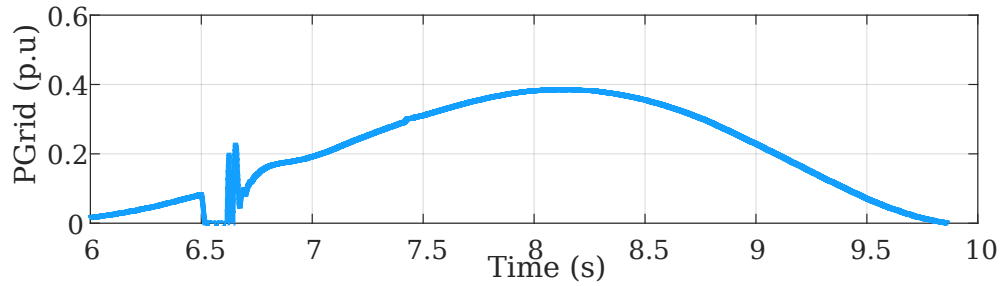
Fig. 20. Impact of the grid SCR on the hybrid PI-TD3 agent performance.



(a)



(b)



(c)

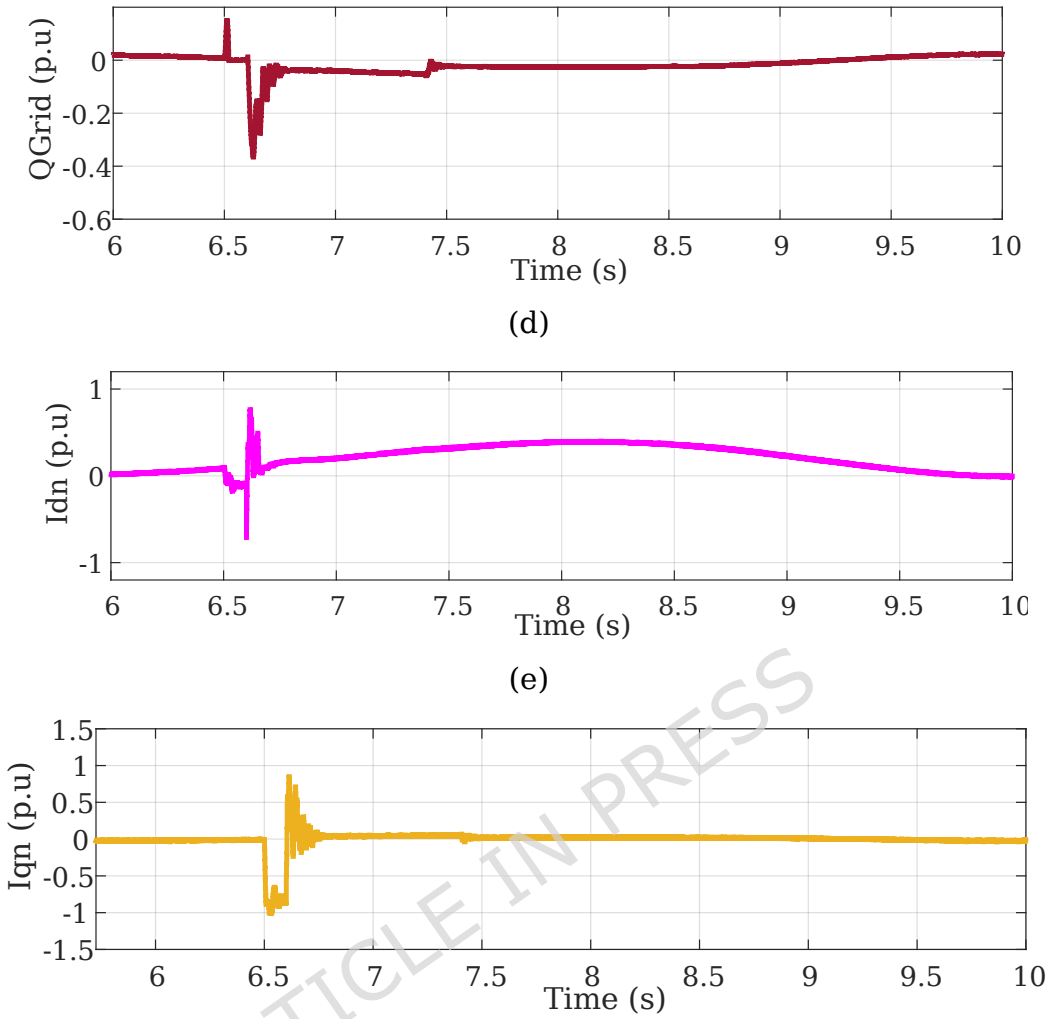


Fig. 21. The system response during an LLLG fault under grid with SCR = 5.

The results emphasize that the system can survive extreme conditions (SCR = 1.5). However, under lower values like SCR =1, the system wouldn't be able to keep V_{PCC} or V_{DC} at 1 pu. Moreover, Fig. 20(d) shows that as the grid becomes weaker (lower SCR), the inverter injects more reactive power to keep V_{PCC} at its nominal value. During a LLLG in a grid with SCR =5, the TD3 controller achieved the control objectives as before. However, the TD3 agent needs to be trained during the transient state in case the controller is installed in a weak grid with SCR lower than 2 to prevent high overshooting in V_{PCC}.

5. Conclusions

This work presents a novel approach using two twin-delayed deep learning agents to improve the performance of a grid-connected WEC. A TD3 agent is deployed on the generator side to reduce stator electrical losses and increase energy production from sea waves. This agent was compared with the classical PI controller configuration and achieved lower ISE for the *dq* currents. Moreover, the agent was subjected to various sea states and to changes in the floater mass to assess its reliability. On the grid side, two TD3-agent methodologies are developed. The first is the replacement of the classical cascaded PI control loop (four PI controllers) with a TD3 agent that takes multiple observations as input and provides the grid-side *dq* voltages as control actions. In the second approach, a hybrid PI-TD3 agent is deployed. In this control strategy, the first two PI controllers that provide the reference grid *dq* currents are kept as they are, and the second pair of PI controllers is exchanged with a TD3 agent. The two grid-side agents and the PI controllers are benchmarked against each other under the steady-state for 50 s. In addition, the three configurations are analyzed during transient conditions under different grid faults to evaluate the agents' performance. After evaluating the results, the most effective control strategy is the hybrid approach, which combines the TD3 agent with PI controllers. This combination provides the greatest performance in both steady and transient states. To further confirm the hybrid approach’s reliability, the controller’s effectiveness was analyzed at different SCR values (1.5, 3, 5). The controller demonstrated its efficiency in steady-state and transient-state conditions (SCR = 5), further emphasizing its superior performance.

Acknowledgment and Funding Declaration

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/33809)

Data Availability Statement

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Rights Retention Statement

Conflicts of Interest

The authors declare no conflict of interest.

Competing Interests

The authors declare no competing interests.

References

[1] B. G. Reguero, I. J. Losada, and F. J. Méndez, “A global wave power resource and its seasonal, interannual and long-term variability,” *Appl. Energy*, vol. 148, pp. 366–380, Jun. 2015, doi: 10.1016/j.apenergy.2015.03.114.

[2] R. Satymov, D. Bogdanov, M. Dadashi, G. Lavidas, and C. Breyer, “Techno-economic assessment of global and regional wave energy resource potentials and profiles in hourly resolution,” *Appl. Energy*, vol. 364, p. 123119, Jun. 2024, doi: 10.1016/j.apenergy.2024.123119.

[3] M. Jama, A. Wahyudie, H. Shareef, and M. E. Haque, “Wave-to-grid hierarchical coordinated control with energy storage and management for point absorber wave energy converters,” *Energy Conversion and Management: X*, vol. 27, p. 101125, Jul. 2025, doi: 10.1016/j.ecmx.2025.101125.

[4] J. C. C. Henriques, L. M. C. Gato, J. M. Lemos, R. P. F. Gomes, and A. F. O. Falcão, “Peak-power control of a grid-integrated oscillating water column wave energy converter,” *Energy*, vol. 109, pp. 378–390, Aug. 2016, doi: 10.1016/j.energy.2016.04.098.

[5] A. Mahdy, S. H. E. Abdel, and H. M. Hasanien, “A novel maximum energy extraction strategy using lithium-ion batteries for Archimedes wave swing wave energy conversion systems,” *Renew. Energy*, vol. 256, no. PI, p. 124664, 2026, doi: 10.1016/j.renene.2025.124664.

[6] H. A. Said, D. García-Violini, and J. V. Ringwood, “Wave-to-grid (W2G) control of a wave energy converter,” *Energy Conversion and Management: X*, vol. 14, p. 100190, May 2022, doi: 10.1016/j.ecmx.2022.100190.

[7] A. Loukriz, A. Zemmit, A. Bendib, and M. Kichen, “Red-Tailed Hawk (RTH) Algorithm for Optimal PV Reconfiguration in Irrigation Systems Under Partial Shading Conditions,” *ITEGAM- Journal of Engineering and Technology for*

Industrial Applications (ITEGAM-JETIA, vol. 12, no. 58, pp. 621–634, 2026, doi: 10.5935/jetia.v12i58.3207.

[8] Y. Benchenina *et al.*, "Advancing green hydrogen production in Algeria with opportunities and challenges for future directions," *Sci. Rep.*, vol. 15, no. 1, p. 5559, Feb. 2025, doi: 10.1038/s41598-025-90336-1.

[9] A. Mahdy, H. M. Hasanien, W. Helmy, R. A. Turky, and S. H. E. Abdel Aleem, "Transient stability improvement of wave energy conversion systems connected to power grid using anti-windup-coot optimization strategy," *Energy*, vol. 245, p. 123321, Apr. 2022, doi: 10.1016/j.energy.2022.123321.

[10] M. Jabari, S. Ekinci, D. Izci, M. Bajaj, V. Blazek, and L. Prokop, "A robust multi-stage FOPI(1+PIDn) controller for precision control of switched reluctance motors under nonlinear and uncertain conditions," *Energy Reports*, vol. 15, p. 109231, Jun. 2026, doi: 10.1016/j.egyr.2026.109231.

[11] S. Ekinci *et al.*, "A novel hyperbolic tangent-based PID controller tuned by the artificial lemming algorithm for nonlinear steam condenser pressure control," *Sci. Rep.*, vol. 16, no. 1, p. 4600, Jan. 2026, doi: 10.1038/s41598-025-34740-7.

[12] Z. M. Ali, A. M. Ahmed, H. M. Hasanien, and S. H. E. A. Aleem, "Optimal Design of Fractional-Order PID Controllers for a Nonlinear AWS Wave Energy Converter Using Hybrid Jellyfish Search and Particle Swarm Optimization," *Fractal and Fractional*, vol. 8, no. 1, p. 6, Dec. 2023, doi: 10.3390/fractalfract8010006.

[13] K. Wang *et al.*, "A method for smoothing grid power fluctuation in hydraulic energy storage type wave energy converter based on active power compensation," *Energy*, vol. 342, p. 139651, Jan. 2026, doi: 10.1016/j.energy.2025.139651.

[14] F. Wu *et al.*, "Modeling, Control Strategy, and Power Conditioning for Direct-Drive Wave Energy Conversion to Operate With Power Grid," *Proceedings of the IEEE*, vol. 101, no. 4, pp. 925–941, Apr. 2013, doi: 10.1109/JPROC.2012.2235811.

[15] S. Rasool, M. R. Islam, K. M. Muttaqi, and D. Sutanto, "An Advanced Control Strategy for a Smooth Integration of Linear Generator Based Wave Energy Conversion System with Distribution Power Grids," in *2019 IEEE Industry Applications Society Annual Meeting, IAS 2019*, IEEE, 2019. doi: 10.1109/IAS.2019.8912314.

[16] A. Parwal *et al.*, "Energy management for a grid-connected wave energy park through a hybrid energy storage system," *Appl. Energy*, vol. 231, pp. 399–411, Dec. 2018, doi: 10.1016/j.apenergy.2018.09.146.

[17] M. I. Marei, M. Mokhtar, and A. A. El-Sattar, “MPPT strategy based on speed control for AWS-based wave energy conversion system,” *Renew. Energy*, vol. 83, pp. 305–317, Nov. 2015, doi: 10.1016/J.RENENE.2015.04.039.

[18] A. Mahdy, H. M. Hasanien, S. H. E. A. Aleem, M. Al-Dhaifallah, A. F. Zobaa, and Z. M. Ali, “State-of-the-Art of the most commonly adopted wave energy conversion systems,” *Ain Shams Engineering Journal*, p. 102322, May 2023, doi: 10.1016/j.asej.2023.102322.

[19] D. Lin, X. Li, S. Ding, H. Wen, Y. Du, and W. Xiao, “Self-Tuning MPPT Scheme Based on Reinforcement Learning and Beta Parameter in Photovoltaic Power Systems,” *IEEE Trans. Power Electron.*, vol. 36, no. 12, pp. 13826–13838, 2021, doi: 10.1109/TPEL.2021.3089707.

[20] A. Rajamallaiah, S. P. K. Karri, M. L. Alghaythi, and M. S. Alshammari, “Deep Reinforcement Learning Based Control of a Grid Connected Inverter with LCL-Filter for Renewable Solar Applications,” *IEEE Access*, vol. 12, no. January, pp. 22278–22295, 2024, doi: 10.1109/ACCESS.2024.3364058.

[21] P. Qashqai, M. Babaie, R. Zgheib, and K. Al-Haddad, “A Model-Free Switching and Control Method for Three-Level Neutral Point Clamped Converter Using Deep Reinforcement Learning,” *IEEE Access*, vol. 11, pp. 105394–105409, 2023, doi: 10.1109/ACCESS.2023.3318264.

[22] G. A. Ghazi *et al.*, “Dandelion Optimizer-Based Reinforcement Learning Techniques for MPPT of Grid-Connected Photovoltaic Systems,” *IEEE Access*, vol. 12, no. March, pp. 42932–42948, 2024, doi: 10.1109/ACCESS.2024.3378749.

[23] S. He, H. Wang, J. Yan, C. Tao, Y. Liu, and S. Han, “A low-computational physics-guided deep learning model for wind farm flow control under time-varying wind conditions,” *Energy*, vol. 332, p. 137048, Sep. 2025, doi: 10.1016/j.energy.2025.137048.

[24] Z. Wu, Y. Li, X. Zhang, S. Zheng, and J. Zhao, “Distributed voltage control for multi-feeder distribution networks considering transmission network voltage fluctuation based on robust deep reinforcement learning,” *Appl. Energy*, vol. 379, p. 124984, Feb. 2025, doi: 10.1016/j.apenergy.2024.124984.

[25] X. Wang, P. Wang, R. Huang, X. Zhu, J. Arroyo, and N. Li, “Safe deep reinforcement learning for building energy management,” *Appl. Energy*, vol. 377, p. 124328, Jan. 2025, doi: 10.1016/j.apenergy.2024.124328.

[26] F. G. Pierart, P. G. Campos, C. E. Basoalto, J. Rohten, and T. Davey, “Experimental Implementation of Reinforcement Learning Applied to Maximise Energy from a Wave Energy Converter,” *Energies (Basel).*, vol. 17, no. 20, pp. 1–13, 2024, doi: 10.3390/en17205087.

[27] Z. M. Ali, A. Mahdy, O. Aldosari, F. Aldawsari, S. H. E. Abdel Aleem, and H. M. Hasanien, “Deep reinforcement learning for energy maximization in AWS wave energy systems with supercapacitor storage unit,” *Ain Shams Engineering Journal*, vol. 17, no. 6, p. 104169, Jun. 2026, doi: 10.1016/j.asej.2026.104169.

[28] H. Su, H. Qin, Z. Wen, H. Liang, and H. Jiang, “Deep reinforcement learning for real-time latching control of wave energy converter in non-predicted irregular wave environments,” *Renew. Energy*, vol. 257, p. 124821, Feb. 2026, doi: 10.1016/j.renene.2025.124821.

[29] E. Anderlini, S. Husain, G. G. Parker, M. Abusara, and G. Thomas, “Towards Real-Time Reinforcement Learning Control of a Wave Energy Converter,” *J. Mar. Sci. Eng.*, vol. 8, no. 11, p. 845, Oct. 2020, doi: 10.3390/jmse8110845.

[30] H. Liang, H. Qin, H. Su, Z. Wen, and L. Mu, “Environmental-Sensing and adaptive optimization of wave energy converter based on deep reinforcement learning and computational fluid dynamics,” *Energy*, vol. 297, p. 131254, Jun. 2024, doi: 10.1016/j.energy.2024.131254.

[31] M. Ye, C. Zhang, Y. Ren, Z. Liu, O. J. Haidn, and X. Hu, “Adaptive optimization of wave energy conversion in oscillatory wave surge converters via SPH simulation and deep reinforcement learning,” *Renew. Energy*, vol. 246, p. 122887, Jun. 2025, doi: 10.1016/J.RENENE.2025.122887.

[32] Paul Gieske, “Model Predictive Control of a Wave Energy Converter: Archimedes Wave Swing,” p. 101, 2007.

[33] M. G. de Sousa Prado, F. Gardner, M. Damen, and H. Polinder, “Modelling and test results of the Archimedes wave swing,” *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, vol. 220, no. 8, pp. 855–868, Dec. 2006, doi: 10.1243/09576509JPE284.

[34] C. L. Bretschneider, “Technical Memorandum No. 118: Wave Variability and Wave Spectra for Wind-Generated Gravity Waves,” Washington, D.C, 1959.

[35] J. R. Morison, J. W. Johnson, and M. P. O’Brien, “Experimental Studies of Forces on Piles,” *Coastal Engineering Proceedings*, vol. 1, no. 4, p. 25, 2000, doi: 10.9753/icce.v4.25.

[36] F. Wu, X. P. Zhang, P. Ju, and M. J. H. Sterling, “Modeling and control of AWS-based wave energy conversion system integrated into power grid,” *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1196–1204, 2008, doi: 10.1109/TPWRS.2008.922530.

[37] S. Fujimoto, H. Van Hoof, and D. Meger, “Addressing Function Approximation Error in Actor-Critic Methods,” *35th International Conference on Machine Learning, ICML 2018*, vol. 4, pp. 2587–2601, 2018.

[38] A. Mahdy, H. M. Hasanien, W. H. A. Hameed, R. A. Turky, S. H. E. A. Aleem, and E. A. Ebrahim, “Nonlinear Modeling and Real-Time Simulation of a Grid-Connected AWS Wave Energy Conversion System,” *IEEE Trans. Sustain. Energy*, vol. 13, no. 3, pp. 1744–1755, Jul. 2022, doi: 10.1109/TSTE.2022.3174176.