IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 April 2022, accepted 6 May 2022, date of publication 12 May 2022, date of current version 18 May 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3174625

Efficient Load Frequency Control of Renewable Integrated Power System: A Twin Delayed DDPG-Based Deep Reinforcement Learning Approach

JUNAID KHALID[1], MAKBUL A.M. RAMLI[1,2], MUHAMMAD SAUD KHAN[1], AND TAUFAL HIDAYAT[1]

[1]Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Center of Research Excellence in Renewable Energy and Power Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Junaid Khalid (jallahrakhadhillon@stu.kau.edu.sa)

This work was supported in part by the Deputyship for Research and Innovation, Ministry of Education of Saudi Arabia, and in part by the King Abdulaziz University, Jeddah, under grant IFPRC-190-135-2020.

ABSTRACT Power systems have been evolving dynamically due to the integration of renewable energy sources, making it more challenging for power grids to control the frequency and tie-line power variations. In this context, this paper proposes an efficient automatic load frequency control of hybrid power system based on deep reinforcement learning. By incorporating intermittent renewable energy sources, variable loads and electric vehicles, the complexity of the interconnected power system is escalated for a more realistic approach. The proposed method tunes the proportional-integral-derivative (PID) controller parameters using an improved twin delayed deep deterministic policy gradient (TD3) based reinforcement learning agent, where a non-negative fully connected layer is added with absolute function to avoid negative gain values. Multi deep reinforcement learning agents are trained to obtain the optimal controller gains for the given two-area interconnected system, and each agent uses the local area control error information to minimize the deviations in frequency and tie-line power. The integral absolute error of area control error is used as a reward function to derive the controller gains. The proposed approach is tested under random load-generation disturbances along with nonlinear generation behaviors. The simulation results demonstrate the superiority of the proposed approach compared to other techniques presented in the literature and show that it can effectively cope with nonlinearities caused by load-generation variations.

INDEX TERMS Load frequency control, deep reinforcement learning, twin delayed deep deterministic policy gradient (TD3), hybrid power system.

I. INTRODUCTION

The growing energy demand, environmental impacts, and depletion of fossil fuels have led to large-scale use of renewable energy sources (RES). The utilization of these RES results in complex and dynamic electric power systems [1]. Therefore, it is becoming more challenging for modern grids to maintain the frequency and tie-line power within a specified limit in interconnected areas. The deviation in frequency causes an imbalance between electric load and the generation [2]. As the load continuously varies and if there would not be an immediate action to mitigate the problem then it could lead to a severe damage. Recently, due to large penetration of intermittent renewable energy sources into the grid led to a total blackout of the power system [3]. Hence, effective control strategies are vital under uncertain conditions in order to achieve a balance between the system reliability and efficiency. Therefore, automatic load frequency control (ALFC) plays an important role in maintaining load-generation balance by regulating the tie-line power flow and frequency oscillations between interconnected areas.

At present, classical proportional integral derivative (PID) type controllers are being used by utilities for load frequency control (LFC) because of their simple structure, high

The associate editor coordinating the review of this manuscript and approving it for publication was M. Mejbaul Haque.

reliability, and better performance-to-cost ratio. PID controller gain values are being tuned over the decades based on experience, utilizing trial-and-error procedures and conventional tuning methods such as Ziegler-Nicholas, but these strategies perform poorly under random load variations and wide range of operating conditions [4]. Over the years, researchers have proposed several intelligent and optimized based control strategies for LFC. Fuzzy logic and adaptive neuro fuzzy inference system (ANFIS) are proposed to tune the PID parameters in [5], [6]. However, a fuzzy system needs field expertise to tune the membership functions and it is difficult to acquire the specific knowledge due to its inadaptability [7]. Recently, many advanced control techniques are proposed for LFC, such as model predictive control (MPC) [8], sliding mode control (SMC) [9], disturbance rejection control [10], and variable structure control [11]. But these controllers are complex and not widely used in the industry, so it is required to improve the PID controller owing to its widespread applications.

As unideal gains are the primary impediment in optimum settings of PID controller, the gain values are therefore derived by heuristic approaches like genetic algorithm (GA) [12], particle swarm optimization (PSO) [13], firefly algorithm (FA) [14], grey wolf optimization (GWO) [15], ant colony optimization (ACO) [16], etc. However, mostly these schemes are only proposed for conventional power systems without considering RES and nonlinear constraints. Apart from that, researchers also have proposed cascade controllers for LFC in articles [17], [18], but these types of techniques required additional controller that also has to be tuned, so it increases the complexity of the strategy.

In recent years, reinforcement learning (RL) based control techniques have been identified as a promising solution for the modern grid. A critical literature review on electric power system control using reinforcement learning has been presented in [19]. Reinforcement learning exhibits superiority over conventional control schemes because of its self-learning approach via an interactive trial and error method based on observations it gets from the dynamic environment. Hence, reinforcement learning can make decisions and solve realistic control problems more effectively. There are some studies proposed in the literature to control the frequency of an interconnected area using reinforcement learning schemes. Data-driven RL based control techniques are presented in [20]–[22] for LFC of multi-area power systems. However, while designing traditional RL agents, the degree of action discretization becomes crucial since control action is taken from a low-dimensional action domain, resulting in limited control performance [23]. Here, deep learning was combined with RL to overcome these deficiencies, which is called deep reinforcement learning. A new approach is proposed in [24] for frequency control using DRL in the continuous action domain, but this kind of technique lacks a constant gradient signal due to the concurrent learning behavior of agents [25].

To solve the continuous control problems, deep deterministic policy gradient (DDPG) was put forward by Lillicrap *et al.* [23] and it does not necessitate the discretization of both the states and actions. Recently, Yan *et al.* [26] have proposed a multi-agent deep reinforcement learning (MA-DRL) approach for multi-area LFC using DDPG. The concept behind that article is an offline centralized learning and online individual application for each control area, where the objective function is maximized by formulating the controller as an MA-DRL problem. However, since DDPG updates the Q-value in the same way as deep Q-networks (DQN) does, it inherits the drawback of overestimation of Q-values, which may lead to suboptimal policy and incremental bias [27]. Moreover, as the authors [26] first implemented the PID controller on the power system to collect the data for initialization of the agent, so specific dataset may lead to sub-optimum convergence under continuous variations of load-generation. A grid-area coordinated LFC technique based on an effective exploration with multi-agent DDPG (EE-MADDPG) is presented in [28], but it cannot be practically implemented on actual grid due to abrupt changing of power grid. Furthermore, as discussed earlier these types of controlling schemes are not widely being used in the industry compared to the PID controller.

Therefore, in this paper, we propose a twin delayed deep deterministic policy gradient (TD3) based deep reinforcement learning approach to fine-tune the PID controller parameters under uncertain conditions. TD3 resolves the defects of DDPG by employing delayed actors update, double critics and actors, and additive clipped noise on control actions. Moreover, unlike the above papers that use DRL for LFC, our proposed technique can directly interact with the power system model to tune the PID gains for actual power grid. Multi-TD3-agents are trained to minimize the frequency and tie-line power deviations of the power system, where each agent uses the local area control error information to decide the action. Furthermore, for better performance we replaced the actor-network's fully connected layer with the new layer consisting of function $y = abs(weights) * x$. This new layer ensures that the weights are positive, as gradient descent optimization may lead the weights to negative values. Moreover, a new integrated hybrid power system architecture is proposed for the interconnected system, which comprises of wind, PV, electric vehicle, hydro and thermal plants. Nonlinearities such as generation dead band (GDB) and generation rate constraints (GRC) are also considered because many of the existing studies ignored these realistic nonlinear behaviors.

Our contributions in this paper can be summarized as follows:

- To the best of our knowledge, this paper is the first work that uses the deep reinforcement learning in continuous control action to optimally tune the PID controller parameters.
- We improve the twin delayed deep deterministic policy gradient based agent to avoid negative PID gain values

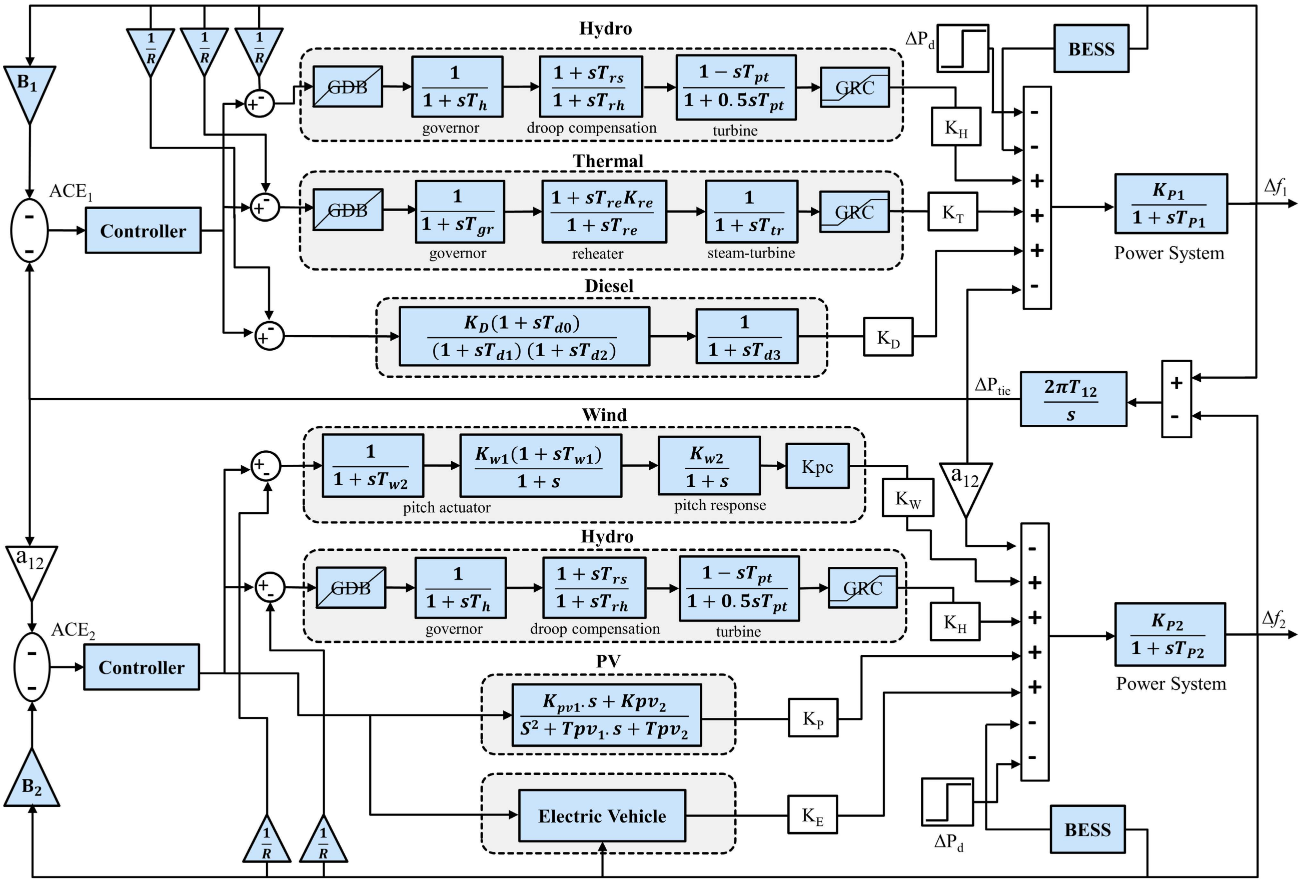

FIGURE 1. Two-area interconnected power system.

while training the agent for LFC, which considerably reduces the computational process.

- We evaluate our novel approach on the given system and compare the performance with metaheuristic and DRL based techniques. In addition, a sensitivity analysis is performed to verify the robustness of the proposed scheme.

The rest of the paper is organized as follows. The modeling of the two-area interconnected system is discussed in Section II. In Section III, the preliminaries used in our work are described. Our proposed TD3 scheme and its implementation is illustrated in Section IV, while the Section V covers the simulation results and discussion of the proposed method. Finally, the article is concluded in Section VI.

II. SYSTEM MODELING

This section briefly discusses the description of the proposed renewable integrated power system. An unequal two-area interconnected power system is under consideration for our study. Area 1 consists of hydro, reheat thermal and diesel, while area 2 integrates the wind, hydro, PV and electric vehicle as shown in Figure 1. The differential equations for load frequency control of two area systems are widely reported in the literature [5]–[16], [29]–[31] and can be given as follows.

$$\Delta\dot{P}_{Gi} = \frac{-1}{T_{Gi}}\Delta P_{Gi} + \frac{-1}{RiT_{Gi}}\Delta f_i + \frac{-1}{T_{Gi}}u_i \tag{1}$$

$$\Delta\dot{P}_{Ti} = \frac{-1}{T_{Ti}}\Delta P_{Gi} + \frac{-1}{T_{Ti}}\Delta P_{Ti} \tag{2}$$

$$\Delta\dot{f}_i = \frac{K_{Pi}}{T_{Pi}}\Delta P_{Ti} + \frac{-1}{T_{Pi}}\Delta f_i - \frac{K_{Pi}}{T_{Pi}}\Delta P_{tie} - \frac{K_{Pi}}{T_{Pi}}\Delta P_{di} \tag{3}$$

where ΔP_{Gi} is the governor position for ith area, ΔP_{Ti} is the power generation level for ith area and Δf_i is the frequency deviation for the ith area. The tie-line connects two areas for power sharing and any particular variation of load in any area can be compensated by the neighboring areas through this tie-line. Mathematically tie-line power can be expressed as

$$P_{12}^0 = \frac{|V_1^0||V_2^0|}{X}sin(\delta_1^0 - \delta_2^0) \tag{4}$$

under any perturbation the tie-line power deviates to

$$\Delta P_{12} = T_{12}(\Delta\delta_1^0 - \Delta\delta_2^0) \tag{5}$$

where T$_{12}$ is

$$T_{12} = \frac{|V_1^0||V_2^0|}{X}cos(\delta_1^0 - \delta_2^0) \tag{6}$$

The final relationship between power angle of machine and frequency deviation will be

$$\Delta P_{12} = \frac{2\pi T_{12}}{S}\left[\Delta f_1(s) - \Delta f_2(s)\right] \tag{7}$$

ACE is the area control error, which is usually the input to the controller that is denoted by

$$ACE_1 = B_1 \Delta f_1 + \Delta P_{tie} \tag{8}$$

$$ACE_2 = B_2 \Delta f_2 + a_{12} \Delta P_{tie} \tag{9}$$

B_1 and B_2 in Figure 1 are the frequency bias parameters that can be described as $B_i = (1/R_i) + D_i$, whereas R_i is the governor speed regulation parameter and D_i is the dependency parameter. The area size ratio is shown as '$a'_{12} = -P_{r1}/P_{r2}$, where P is the power capacity (MW) for each area. The detailed information of the parameters used for simulation is listed in the Appendix and taken from [29]–[31]. The block diagrams and transfer functions of the conventional power system are extensively discussed in the literature [5]–[16], whereas we have integrated the electric vehicle (EV), wind and PV into the system, and their details are given below.

A. ELECTRIC VEHICLE MODEL

An aggregate model of the EV comprised of a battery charger and primary frequency control is illustrated in Figure 2. EV fleets can compensate the unscheduled load by exchanging power between battery and the grid via a charger. The dead band function along with droop characteristics is taken into account since there is a possibility that all EVs may disconnect from the grid resulting in frequency deviation. The upper and lower limits of the dead band are set to 10 mHz and -10 mHz respectively, whereas the droop coefficient (R_{ev}) value is taken same as other plants. K_{EV} represents the EV gain and the value of K_{EV} (between 0-1) determined by the EVs' state of charge (SOC), while the battery time constant is represented by T_{EV}. ΔP_{AG}^{max} and ΔP_{AG}^{min} are the maximum and minimum power outputs of the EV fleets and these reserves can be calculated as follows [32].

$$\Delta P_{AG}^{max} = +\left[\frac{1}{N_{EV}} \times (\Delta P_{EVi})\right] \tag{10}$$

$$\Delta P_{AG}^{min} = -\left[\frac{1}{N_{EV}} \times (\Delta P_{EVi})\right] \tag{11}$$

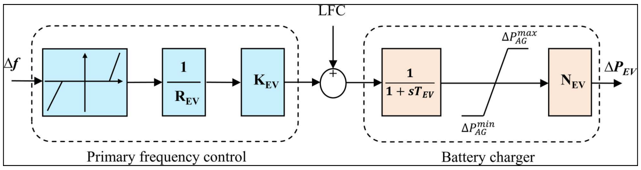

FIGURE 2. Block diagram of electric vehicle model.

The incremental generation change of EV in the area is denoted by P_{EV}. N_{EV} indicates the total number of electric vehicles connected to the system.

B. WIND GENERATION MODEL

A wind turbine (WT) based on a doubly-fed induction generator (DFIG) is investigated in this study. Wind turbines convert wind energy into electricity and the output power can be characterized as follows.

$$P_W = 0.5 \rho A C_P V_W^3 \tag{12}$$

Here ρ, A, C_p and V represents the air density, blade swept area, power coefficient and wind speed respectively. The wind turbine power coefficient is:

$$C_p(\lambda, \beta) = (0.44 - 0.0167\beta) sin\left(\frac{\pi(\lambda - 2)}{15 - 0.3\beta}\right) - 0.00184(\lambda – 3) \tag{13}$$

where β is the blade pitch angle, and λ is the tip speed ratio. The transfer function can be written as [33]

$$G_w = \frac{K_{w1}(1 + sT_{w1})K_{w2}}{(1 + sT_{w2})(1 + s)(1 + s)} \tag{14}$$

A wind energy system can cause instability in the system due to its intermittent nature; hence continuous power fluctuation can be handled by a battery. Battery energy storage system (BESS) stores excessive electrical energy and if equipped with a large battery bank, it can offer a great amount of power supply for a longer length of time. The simplified transfer function of the BESS is expressed as follows.

$$G_{BESS} = \frac{K_{BESS}}{(1 + sT_{BESS})} \tag{15}$$

C. PHOTOVOLTAIC MODEL

Photovoltaic (PV) modules are solar energy-generating components. The relationship between voltage and current is nonlinear because of the variation in solar radiations throughout the day. Therefore, to increase the output power of the PV panel, a maximum power point tracker (MPPT) must be used. The following is a description of the PV plant's transfer function with MPPT [34].

$$G_{PV} = \frac{sK_{PV1} + K_{PV2}}{(s^2 + sT_{PV1} + T_{PV2})} \tag{16}$$

K_{PVi} and T_{PVi} represent the gains and time constants of the PV system respectively. Incremental conductance (IC) method is used to extract MPP from the PV system under the following conditions.

$$\begin{cases} \dfrac{dP_{PV}}{dV_{PV}} > & 0 \; at \; right \\ \dfrac{dP_{PV}}{dV_{pv}} = & 0 \; at \; MPP \\ \dfrac{dP_{PV}}{dV_{PV}} < & 0 \; at \; left \end{cases} \tag{17}$$

D. NONLINEAR GENERATION BEHAVIORS

Generation rate constraint (GRC) and dead band (GDB) are incorporated into the system for a more realistic approach. Power generation can only vary at a certain limit called GRC, on the other hand, GDB is the steady-state speed change until the governor valves position changes. GDB has a significant effect that may lead to random fluctuation, and the factors that contribute to it are backlash in different governor linkages between servo piston and camshaft, and valve overlapping in hydraulic relays [35]. The nonlinear models are shown in Figure 3, and for thermal plant the values of GRC and GDB are taken as ±3% /min and 0.06% (0.036 Hz), respectively. The GRC lowering and raising values of hydropower plants are 360%/min and 270%/min, respectively, whereas GDB limit is 0.02%. Thermal unit's GDB is incorporated in the governor transfer function as given below

$$G_T = \frac{N_1 + (N_2/w_0)s}{1 + sT_T} \tag{18}$$

where N_1, N_2 and w_0 are computed as 0.8, -0.2 and π respectively.

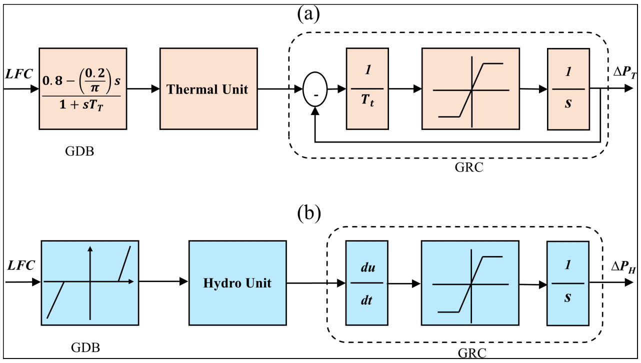

FIGURE 3. GRC and GDB model for (a) Thermal and (b) Hydro.

Moreover, the area participation factor (apf) for each plant is considered to determine how much each unit will contribute to the nominal loading. K_H, K_T, and K_D of area 1 are the apfs of hydro, thermal, and diesel plants, respectively. Similarly, the apfs for area 2 are specified, where sum of these factors must be equal to 1 for each area. The apfs for each unit are listed in the Appendix.

III. PRELIMINARIES

In this section, a brief description of the deep reinforcement learning techniques that will be used in our study are presented.

A. DEEP DETERMINISTIC POLICY GRADIENT

DDPG is an improved class of deterministic policy gradient that combines DPG and DQN, and is a model free off-policy actor-critic algorithm. Moreover, it can be used in continuous space using policy-function (actor) and Q-function (critic) framework, which is essential for analysis of the power system as it operates in continuous action because of varying load and generation. A general network of actor and critic is shown in Figure 4. The critic uses temporal difference (TD) technique to update its parameters in the same way as DQN does, whereas DPG algorithm is used to update the actor via $\alpha = \mu(s|\theta_\mu) + N$, here N represents random noise function.

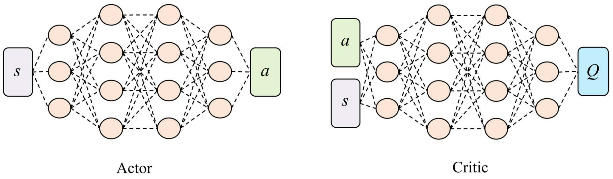

FIGURE 4. Actor and critic networks.

Exponential smoothing is used to update the corresponding θ_μ and θ_Q parameters of the actor and critic network as stated below [36].

$$\begin{aligned} \theta_{\mu'} &= \tau\theta_\mu + (1-\tau)\theta_{\mu'} \quad (actor) \\ \theta_{Q'} &= \tau\theta_Q + (1-\tau)\theta_{Q'} \quad (critic) \end{aligned} \tag{19}$$

The learning stability may be improved owing to slow and smooth variations of the target network and hyperparameters. Using critic framework, the action values can be estimated with the Bellman equation.

$$Q'(s,a) = E\left[r(s,a) + \gamma Q'(s',a')\right] \tag{20}$$

Next, $y = r + \gamma Q'(s',a')$ is used as a TD-error with a discounting factor $\gamma \ll 1$, and to update the critic parameters minimize the loss function across all samples.

$$L = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(s_i,a_i))^2 \tag{21}$$

For training, DDPG employs the experience relay (ER) technique, in which a random dataset is selected from the reply buffer and trained in a mini-batch scheme. Through mapping the state of the provided action, the current network's actor parameters are updated via action value function and then updated them using the neural network gradient backpropagation. To maximize the expected discounted reward, the following policy gradient is used [23].

$$\nabla_{\theta_\mu} J \approx \frac{1}{M}\sum_{i=1}^{M}\left[\nabla_a Q(s,a)|_{s=s_i,a=\mu(s_i|\theta_\mu)} \nabla_{\theta_\mu}\mu(s|\theta_\mu)|_{s_i}\right] \tag{22}$$

To learn the parameterized policy, the Actor-Critic technique converts Monte Carlo based updates into TD. Meanwhile, classic on-policy is transformed to off-policy by adding experience replay from DQN and a target network, which enhances sample efficiency.

B. TWIN-DELAYED DEEP DETERMINISTIC POLICY GRADIENT (TD3)

The performance of Q-learning method is known to be affected by overestimation of the value function, so the policy update will be negatively affected if the overestimation persists throughout training. Because of these limitations, approaches such as double Q-learning and double DQN are developed, which employ two value networks to separate

Q-value and actions' selection updates. Twin delayed DDPG (TD3) [37] solves the overestimation of Q-value using the following three techniques.

To begin, the concept of double Q-learning is imitated by the TD3, which computes the next state value by creating two Q-value networks as given below.

$$\begin{aligned} y_1 &= r + \gamma Q_{\theta'_1}(s', \mu'(s'|\theta_{\mu'})) \\ y_2 &= r + \gamma Q_{\theta'_2}(s', \mu'(s'|\theta_{\mu'})) \end{aligned} \tag{23}$$

To compensate the overestimation of Q-value, the target Q-value is taken as the clipped minimum of two values and then put into the Bellman equation to compute the loss function (same as stated in Eq. 21) and the TD-error as shown in Figure 5 and given below [38].

$$y = r + \gamma^{min}_{i=1,2} Q'_i(s', a') \tag{24}$$

Even though this Q-value update rule may result in an underestimating bias when compared to the classic Q-learning technique, the action values will not be openly passed on via policy update.

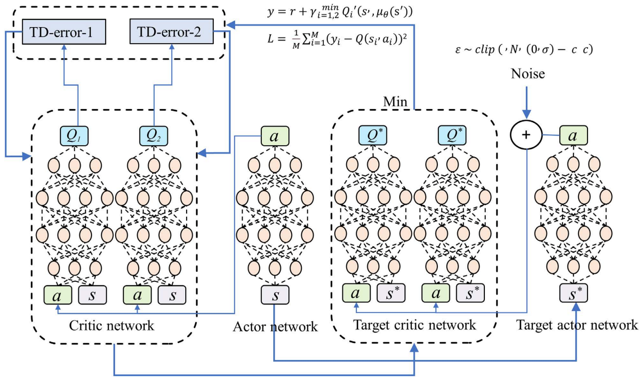
$y = r + \gamma^{min}_{i=1,2} Q_i'(s', \mu_\theta(s'))$
$L = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(s_i, a_i))^2$
$\varepsilon \sim clip(, N, (0, \sigma) - c\ c)$
Noise
TD-error-1
TD-error-2
Min
Q_1
Q_2
a
Q^*
Q^*
+
a
a
s
a
s
s
a
s^*
a
s^*
s^*
Critic network
Actor network
Target critic network
Target actor network

FIGURE 5. Architecture of Twin-delayed DDPG (TD3).

Moreover, to achieve better convergence the target network is set up being a deep function approximator that offers constant objectives while learning phase. On the other hand, the observed states are sensitive to divergence if the error is integrated. Therefore, compared to the value network the policy network is intended to update at a lower frequency in order to limit the error propagation, hence a high-quality update can be obtained.

Finally, to avoid overfitting, the Q-value computation needs to be smoothed in order to resolve the trade-off between bias and variability. Hence, for each action a clipped normal distribution noise is applied as a regularization, resulting in the revised target update as shown below [37].

$$\begin{aligned} y &= r + \gamma Q_\theta(s', \mu'(s'|\theta_{\mu'}) + \varepsilon) \\ \varepsilon &\sim clip(N(0, \sigma), -c, c) \end{aligned} \tag{25}$$

IV. PROPOSED METHOD

The twin delayed DDPG-based agent is trained to act as an LFC controller to optimally tune the PID parameters. Multi-agents have been trained where each area has its own frequency controller (agent) in the proposed interconnected system, and the elements involved in this formulation are stated as follows.

A. ENVIRONMENT

Everything in an interconnected power system apart from the agent is referred to as an environment. An agent takes the environment's state as information at every time step to choose an appropriate action, and then the environment gives back a reward and new state against that particular action.

B. OBSERVATIONS

The frequency response that will be used by the TD3 algorithm, policy, and reward function is represented by the state or observations.

C. REWARD

To evaluate the agent's behavior against each state the environment gives the feedback to determine whether or not the system is converging its objectives. As a result, reward function directly influences the agent to take actions that maximize the values in order to approach objective function.

D. ACTION

It is the agent output in the form of a control signal to the power plant and its value decided by the policy to maximize reward at a certain state.

The implementation of the twin delayed DDPG agent is illustrated in Figure 6. The area control error (ACE) is the input/state for each agent in the proposed interconnected system. The state (s) or observations are given as proportional, integral, and derivative of the ACE to calculate the action (a) of each agent in both areas. Based on reward function and frequency response the agent tries out different PID values to interact with the power system and this action exploration remains continuous until its approach specified objectives. To get the optimal PID parameters the reward function plays an important role in effectively taking the actions toward solving the defined load frequency control problem because reward affects the Bellman equation (Eq. 24) of the proposed TD3 algorithm. As the agent learns all by itself by continuously updating its parameters, so the proper reward function will help in fast convergence with less computation and high performance. In this paper, the absolute sum of frequency deviation and tie-line power is defined as the objective/reward function to minimize the fluctuation of tie-line power and frequency in both areas. The reward function is stated as follows.

$$R = -\sum_{i=1}^{T} \left[\left| B_i \Delta f_i \right| + |\Delta P_{tie}| \right] \tag{26}$$

where negative will minimize the error, thus maximizes the reward. If a particular agent's action is not taken towards minimizing the error, then a penalty will be applied which

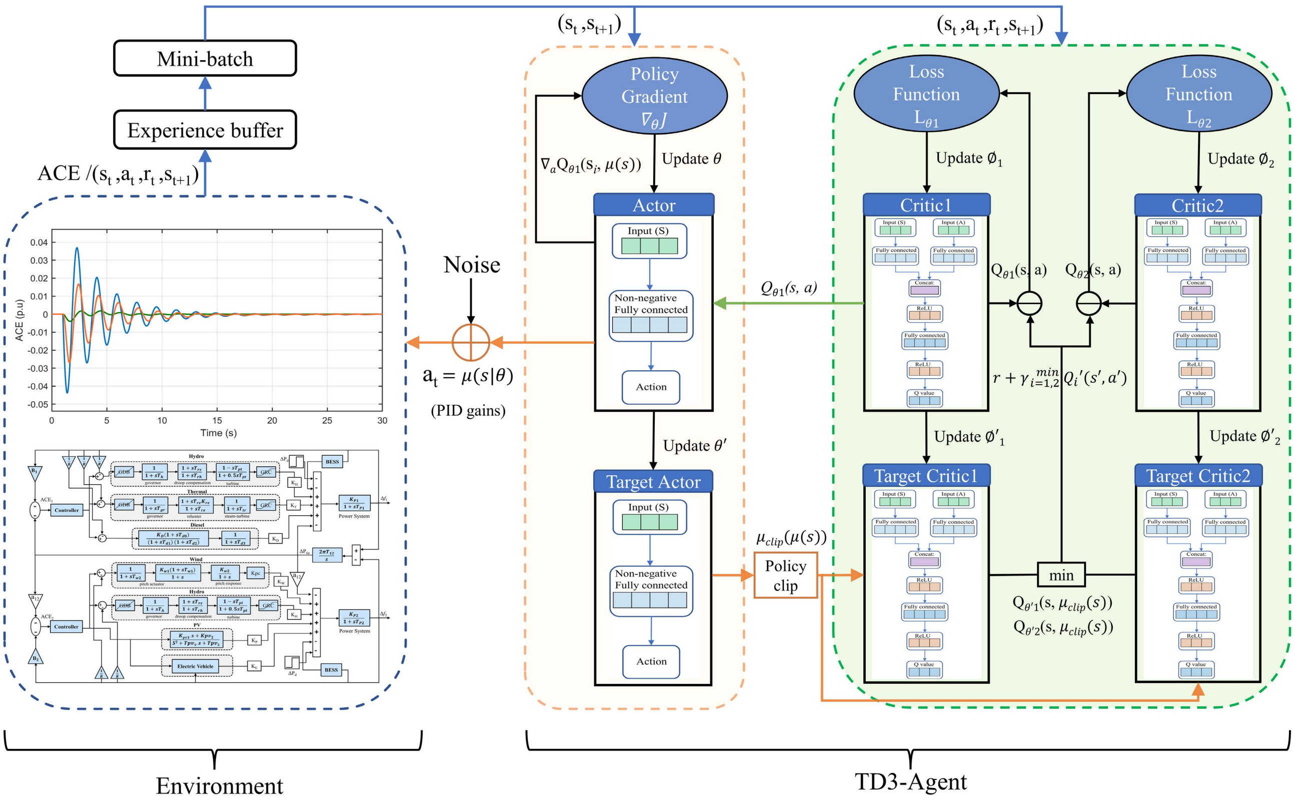

FIGURE 6. Workflow of the proposed TD3 approach.

will reduce the reward, so the RL agent keeps exploring the action/PID values that will maximize the reward.

E. DESIGN OF TD3-BASED CONTROLLER

As the primary goal of the scheme is to minimize the frequency and tie-line power under uncertain conditions, the dynamic environment is created by integrating RES and selecting random step disturbance in the power system to train the TD3 agent. The agent receives frequency response from the environment in the form of proportional, integral, and derivative of ACE, and gives the control signal to the environment as an output. The agent consists of critic and actor networks where an actor is known as a policy structure to decide action and the critic is estimated value function. To create the TD3 agent the actor and critic are created as deep neural networks. In actor, we mimic the neural network as PID controller where the feature-input layer is with proportional, integral and derivative of ACE as input and fully connected layer as controller output. Furthermore, we improved the TD3 agent by replacing fully connected layer in actor-network with the new layer that consists of function $y = abs(weights) * x$. This new layer ensures that the weights (PID gains) are positive, as gradient descent optimization may lead the weights to negative values. The parameters which are considered while creating actor and critic networks for the TD3 agent are listed in Table 1.

TABLE 1. Hyperparameters of proposed TD3.

Hyperparameters	Values
Experience buffer length	1000000
Minibatch size	128
Number of steps in an episode	300
Policy update frequency	2
Policy smoothing noise	0.1
Exploration noise	0.1
Target smooth factor	0.005
Discount factor	0.99
Window length of each episode (time)	100
Actor learning rate	0.001
Critic learning rate	0.001
Gradient threshold	1
Optimizer	Adam
Fully connected layer size	32

The critic network shown in Figure 6 is made up of total 9 layers, as it receives the frequency response (s) and actor's action (a), so feature-input layers are used for both inputs. Then, a concatenation layer is added to link both inputs followed by fully connected layers for each input. Rectified linear unit (ReLU) is used between each fully connected layer as an activation function. Adam optimizer is applied to update the parameters of actor and critic networks, while the glorot is used as weights initializer for fully connected layers. To formulate the TD3 agent two critic networks ($Q_1(s,a), Q_2(s,a)$) are created and these two networks help the agent to estimate long-term reward based on states and actions. The structures and parameters of the target actor and target critics are taken similar to the actor-critic. The target actor and target critics parameters are continuously being updated by the agent to improve the optimization's stability. The steps used in implementing the proposed TD3 algorithm for LFC are briefly discussed as follows:

Step 1. Create critic and actor functions for the agent to estimate the value function and policy during training at each time step.

$$\text{critic } Q\,(s, s\,|\,\emptyset)\,,\, Q_t\,(s, s\,|\,\emptyset_t)$$
$$\text{actor } \mu\,(s\,|\,\theta)\,,\, \mu_t\,(s\,|\,\theta_t)$$

Step 2. Specify the agent options such as experience replay buffer length, mini-batch size, and Gaussian noise.

Step 3. Based on specified parameters in step 1 and step 2, create the TD3 agents for both areas.

Step 4. To train the TD3 agent the following algorithm is used.

Once the training is completed the actor network's absolute weights are fetched as the proportional, integral, and derivative gains of the PID controller. The flowchart in Figure 7 illustrates simplified representation for PID tuning.

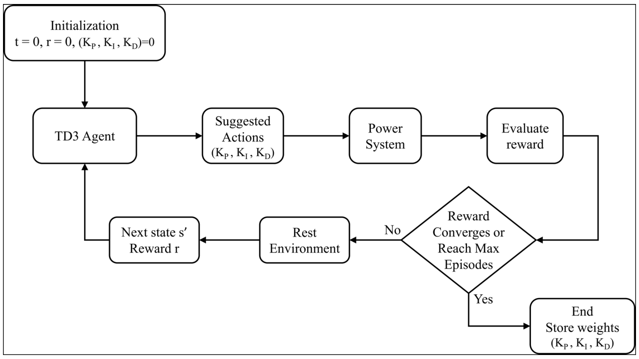
Initialization
t = 0, r = 0, (K_P, K_I, K_D)=0
TD3 Agent
Suggested Actions (K_P, K_I, K_D)
Power System
Evaluate reward
Reward Converges or Reach Max Episodes
No
Yes
Rest Environment
Next state s' Reward r
End
Store weights (K_P, K_I, K_D)

FIGURE 7. Flowchart of PID tuning using proposed TD3.

Algorithm 1 Twin Delayed DDPG.

```
Initialize actor $\mu\,(s\,|\,\theta)$ and critics $Q\,(s, s\,|\,\emptyset)$ networks
Initialize target actor $\mu_t\,(s\,|\,\theta_t)$ and target critics
  $Q_t\,(s, s\,|\,\emptyset_t)$ using primary actor-critic networks'
  parameters
for each episode = 1,..., $M$ do
    Simulate the environment with random
     load-generation disturbance
    Observe the current state as [ACE(s), ACE(s)/s,
     ACE(s)/s+1] and store in experience buffer
    Initialize random exploration noise (Gaussian) $N_t$
    for $t$= 1,..., $T$ do
        Choose an action $a_t = \mu(s|\theta) + N_t$ based on current
         observations/state
        Execute the action and get the details of the reward
         $r_t$ and next state $s_{t+1}$
        Store values $(s_t, a_t, r_t, s_{t+1})$ in experience
         replay buffer
        Sample the random minibatch of store values from
         replay buffer
        Put $y_i = \begin{cases} r_i\ IF\ \ s_{t+1}\ is\ terminalstate \\ r_i + \gamma^{min}_{i=1,2} Q'_i(s', a')\ otherwise \end{cases}$
        Update each critic parameters by minimizing the
         loss function stated in Eq. 21.
        Update actor parameters using
         $\nabla_\theta J \approx \frac{1}{M} \sum_{i=1}^{M} \left[\nabla_a Q(s,a)|_{s=s_i, a=\mu(s_i|\theta)} \nabla_\theta \mu(s|\theta)|_{s_i}\right]$
        Update target actor and target critic parameters
         using smoothing factor
              $\theta_{\mu'} = \tau\theta_\mu + (1-\tau)\,\theta_{\mu'}$
              $\theta_{Q'} = \tau\theta_Q + (1-\tau)\,\theta_{Q'}$
    end for
end for
```

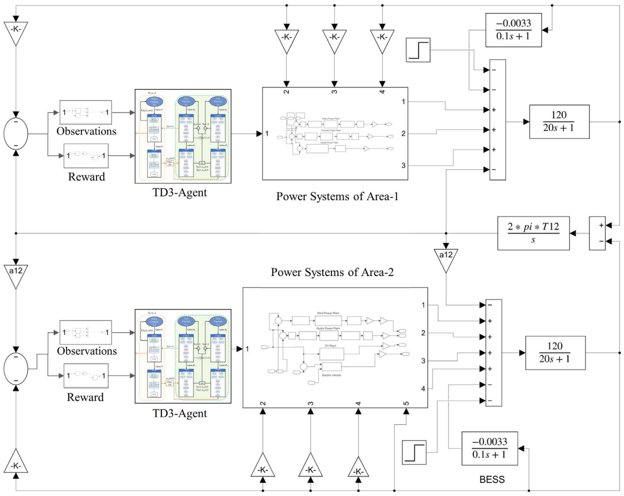

FIGURE 8. Multi-agent TD3 implementation for LFC.

V. RESULTS AND DISCUSSION

The two-area interconnected system shown in Figure 1 is developed in MATLAB/Simulink and the TD3 agent is implemented as a controller for each area in the system to get the optimal PID gains. The configuration of the proposed scheme is illustrated in Figure 8. During training, the algorithm runs the simulation for every episode and the simulation for a single episode continues until it reaches the window's length or triggered the threshold limit. After every episode, each agent gets the reward based on control action. The reward's performance while training both agents is shown

TABLE 2. Pid controller gains for both areas.

Algorithm	Area 1				Area 2			
	K_p	K_i	K_d	IAE	K_p	K_i	K_d	IAE
Proposed TD3	1.4917	1.5085	2.6746	0.0005287	3.0948	3.6408	4.9748	0.0003047
DDPG	0.8927	1.1085	2.0746	0.001206	1.3781	1.9007	1.9846	0.0006915
PSO	1.1917	1.2405	4.2097	0.001099	2.1911	5	2.2046	0.0005977
GA	1.0917	1.7083	5	0.001398	1.2873	2.0872	1.8046	0.0009787

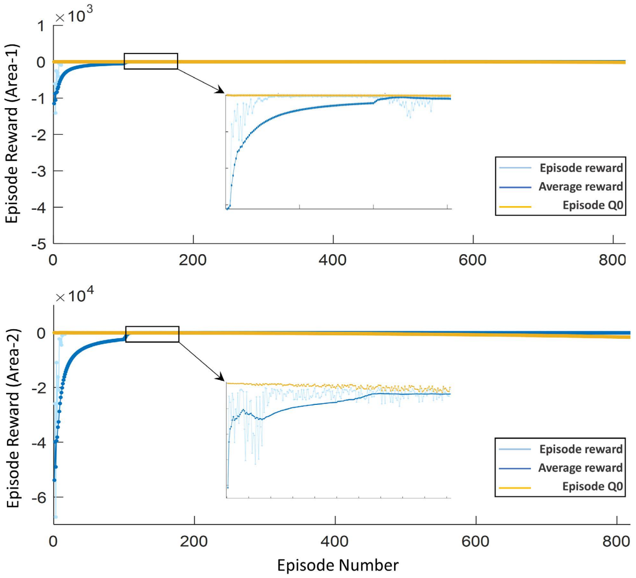

FIGURE 9. Moving average of rewards while training.

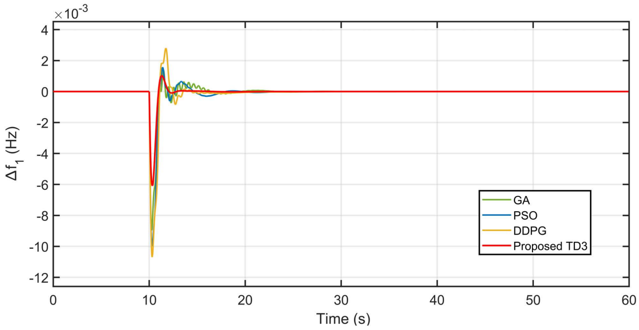

FIGURE 10. Frequency deviation of Area 1 with 1% SLP.

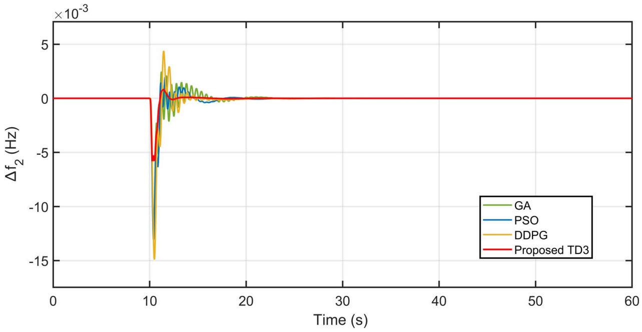

FIGURE 11. Frequency deviation of Area 2 with 1% SLP.

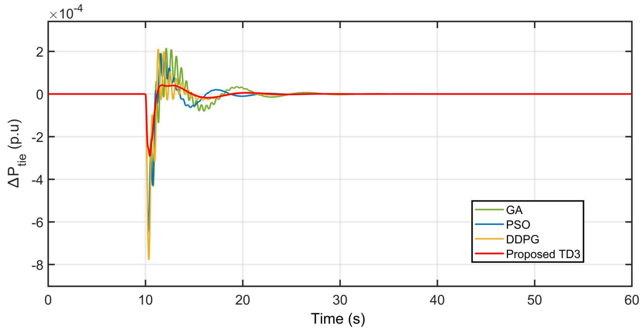

FIGURE 12. Tie-line power deviation with 1% SLP.

in Figure 9. We have given [0 0 0] as initial PID gains to initialize the model, therefore starting episodes received high negative rewards as an error penalty. Area-2 is more heavily penalized due to the presence of RES. The agent tries to maximize the reward by choosing optimal PID gains as control actions. As shown in the figure the model performing better after 200 episodes but to get better results and converge the system at optimal solution 800 episodes are carried out. After training the model, the robustness of the proposed scheme is tested under different scenarios and the results are compared with conventional meta-heuristic and DRL techniques. Table 2 shows the obtained PID controller gains across each algorithm which are taken into consideration for comparison against the proposed approach. While training the model, the lower limit [0 0 0] and the upper limit [5 5 5] of the PID gains are set for every algorithm for a fair comparison. The IAE values in Table 2 show that the proposed TD3 approach gives the minimum error among all listed algorithms.

The random output power fluctuations of RES are provided for system analysis. The power of EV fleet is shown in Figure A1, which illustrates the charging and discharging states of EVs to compensate the unscheduled load by exchanging power between battery and the grid via a controlling charger. To compare the control performance, firstly 1% step load perturbation (SLP) is applied in area 1. The results shown in the Figure 10 clearly indicate the superiority of the proposed TD3 approach, where -0.006 Hz and 0.0011 Hz

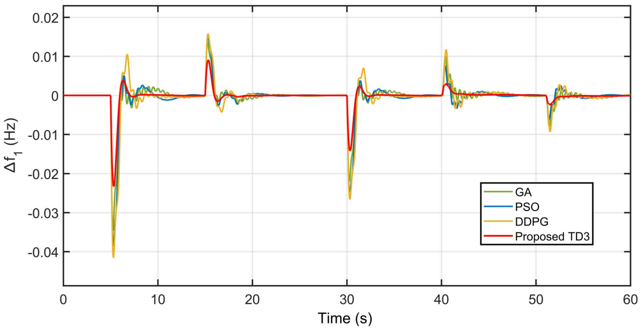

FIGURE 13. Frequency deviation of Area 1 with random SLP.

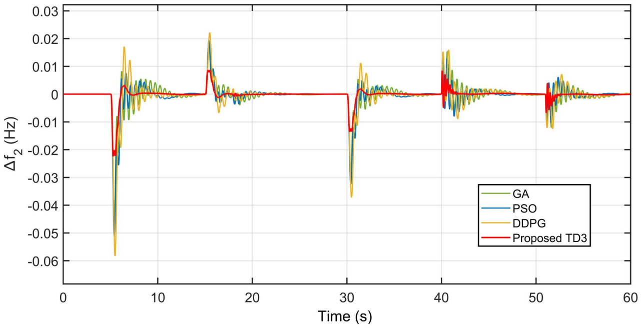

FIGURE 14. Frequency deviation of Area 2 with random SLP.

are the under-shoot (US) and over-shoot (OS) frequency responses of the system The frequency settling time (Ts) is 7.5s compared to the other techniques, which require more than 14s to stabilize the response. The DDPG's US and OS are -0.011 Hz and 0.003 Hz, respectively. The PSO and GA provide nearly identical findings, with a minor variation in frequency responses, where -0.01 Hz and 0.0017 Hz are US and OS of the PSO compared to the GA's -0.009 Hz and 0.0017 Hz, respectively. Moreover, the proposed TD3 scheme efficiently compensated the oscillations while stabilizing the frequency deviations. The detailed LFC performance comparison of all the considered techniques is demonstrated in Table 3.

Furthermore, the robustness of the proposed approach is tested under random step load disturbances in both areas as shown in Figure A2. The responses of frequencies and tie-line power are shown in Figures 13-15. The proposed TD3 method performed better compared to other techniques in terms of minimum undershoot, overshoot and settling time. The GA, PSO, and DDPG exhibit poor performance when the random load disturbance is applied in area 2. The maximum US, OS and Ts for TD3 under random SLP in area 2 is -0.023 Hz, 0.008 Hz and 9s, respectively. For DDPG, the values are -0.059 Hz, 0.022 Hz and 13.5s, respectively. While the PSO's US, OS and Ts are -0.05 Hz, 0.018 Hz and 14s, respectively. The GA performed poorly under random step load disturbance as shown in Figures 13-15, therefore we only considered PSO in further results owing to its slightly better performance than GA. Finally, the performances of all three techniques are assessed under continuous load-generation variations. The given results in Figures 16-18 indicate that the proposed TD3 approach to optimally tune the PID controller parameters outperforms other techniques and provides minimum under-shoot and over-shoot deviation with less oscillations.

The maximum and minimum frequency deviation values in both areas for the proposed TD3 is not crossed 60.05 Hz and 59.95 Hz, respectively. However, for DDPG and PSO the maximum and minimum frequency deviations are 60.19 Hz, 59.83 Hz and 60.15 Hz, 59.84 Hz, respectively. The tie-line power deviates ±0.002 (p.u) in the case of proposed TD3 approach, while for DDPG

TABLE 3. Comparative performance analysis.

Cases	Maximum deviation		Proposed TD3	DDPG	PSO [13]	GA [12]
1% SLP	Δf_1	U_s	-0.006	-0.011	-0.010	-0.009
		O_s	0.001	0.003	0.0017	0.0017
		T_s	7.5s	12s	13s	14s
	Δf_2	U_s	-0.006	-0.015	-0.013	-0.012
		O_s	0.0015	0.004	0.002	0.003
		T_s	8s	12.5s	13s	14.5s
	ΔP_t	U_s	-0.0003	-0.0008	-0.0007	-0.0007
		O_s	0.0001	0.0002	0.0002	0.0002
		T_s	10s	15s	16s	17s
Random SLP	Δf_1	U_s	-0.024	-0.041	-0.036	-0.038
		O_s	0.006	0.016	0.015	0.015
		T_s	9s	13s	15s	15.5s
	Δf_2	U_s	-0.023	-0.059	-0.049	-0.051
		O_s	0.008	0.022	0.018	0.019
		T_s	9s	13.5s	14s	14s
	ΔP_t	U_s	-0.0011	-0.0031	-0.0254	-0.0255
		O_s	0.0005	0.0013	0.0010	0.0011
		T_s	11s	15.5s	n/a	n/a
Dynamic	Δf_1	U_s	-0.05	-0.12	-0.13	-0.14
		O_s	0.05	0.12	0.13	0.14
	Δf_2	U_s	-0.05	-0.17	-0.16	-0.18
		O_s	0.05	0.19	0.15	0.17
	ΔP_t	U_s	-0.0025	-0.008	-0.007	-0.008
		O_s	0.0027	0.01	0.008	-0.01

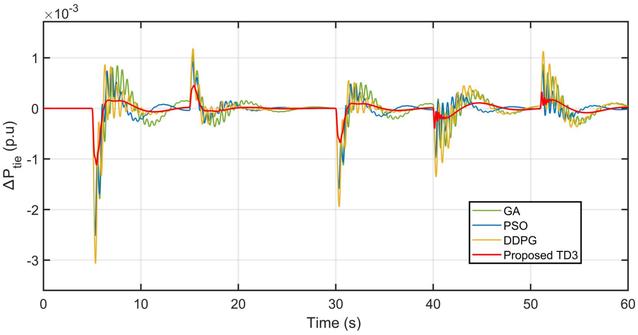

FIGURE 15. Tie-line Power deviation with random SLP.

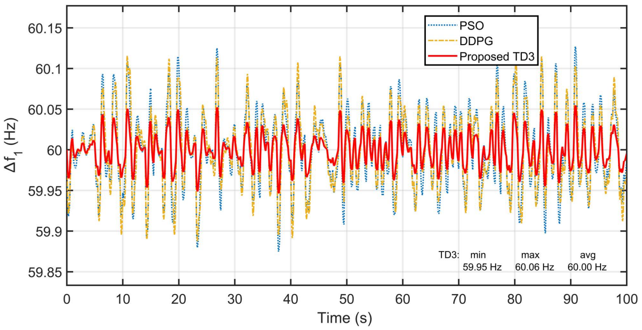

FIGURE 16. Dynamic frequency response of Area 1.

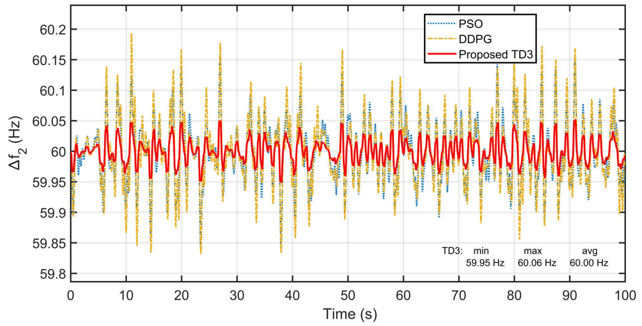

FIGURE 17. Dynamic frequency response of Area 2.

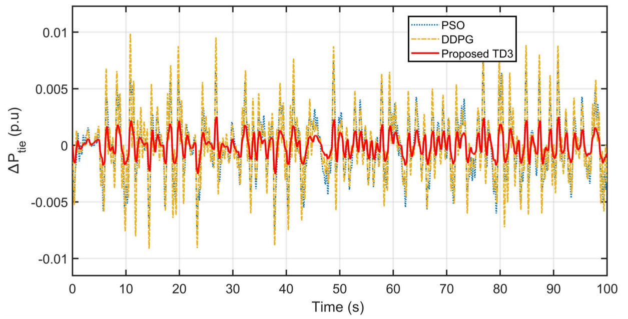

FIGURE 18. Dynamic tie-line power response.

and PSO the deviation is ±0.01 (p.u) and ±0.008 (p.u), respectively. Hence, these results verify the superiority of the proposed TD3 approach against fluctuations of the renewable integrated energy sources into the system.

A. SENSITIVITY ANALYSIS

In this subsection, a sensitivity analysis is carried out to illustrate the robustness of the proposed approach by varying the system parameters and system operating conditions. Since changing the conditions may lead to severe disturbance, the controller parameters should be robust enough to tolerate these changes. To test the proposed approach, parameters such as time constants (T_h, T_{gr}, T_w), gain constants (K_{PV}, K_{EV}), R, and coefficient D are varied in the range of $\mp 50\%$ from nominal values. The optimal PID gains obtained at nominal operating conditions are used to evaluate the performance while varying the system parameters. For sensitivity analysis, only one parameter at a time is changed to 25% while the other parameters are kept at nominal values. The Table 4 shows the control performance of Δf_1, Δf_2, and ΔP_{tie} in steps of 25% parametric variations. The comparison of all listed responses with nominal values reveals the robustness of the proposed approach against system parameters variations, where frequency and tie-line responses are almost overlapping with minor differences.

TABLE 4. Sensitivity analysis with proposed TD3.

Parameters	% of change	Maximum deviation			IAE
		Δf_1	Δf_2	ΔP_{tie}	
Nominal	0	0.0495	0.0451	0.0174	0.0837
T_h	-50%	0.0560	0.0574	0.0266	0.0869
	-25%	0.0524	0.0552	0.0247	0.0847
	+25%	0.0483	0.0431	0.0175	0.0820
	+50%	0.0452	0.0411	0.0157	0.0829
T_{gr}	-50%	0.0527	0.0584	0.0242	0.0878
	-25%	0.0565	0.0537	0.0281	0.0857
	+25%	0.0487	0.3768	0.0163	0.0826
	+50%	0.0563	0.3502	0.0194	0.0801
T_w	-50%	0.0467	0.0359	0.0201	0.0808
	-25%	0.0481	0.0430	0.0190	0.0826
	+25%	0.0501	0.0492	0.0209	0.0846
	+50%	0.0505	0.0550	0.0224	0.0837
K_{PV}	-50%	0.0327	0.0484	0.0242	0.0978
	-25%	0.0365	0.0537	0.0281	0.0957
	+25%	0.0487	0.3768	0.0168	0.0826
	+50%	0.0517	0.3502	0.0195	0.0801
K_{EV}	-50%	0.0576	0.0574	0.0263	0.0863
	-25%	0.0503	0.0582	0.0252	0.0893
	+25%	0.0454	0.3834	0.0146	0.0803
	+50%	0.0565	0.3495	0.0132	0.0800
D	-50%	0.0394	0.0463	0.0276	0.0962
	-25%	0.0365	0.0592	0.0263	0.0937
	+25%	0.0454	0.3539	0.0193	0.0835
	+50%	0.0592	0.3925	0.0139	0.0862
R	-50%	0.0367	0.0467	0.0022	0.0952
	-25%	0.0389	0.0445	0.0021	0.0920
	+25%	0.0445	0.0389	0.0019	0.0871
	+50%	0.0467	0.0367	0.0018	0.0854

The parametric variation responses with T_W and R are illustrated in Figures 19 and 20 respectively, which confirm the robustness of the proposed TD3 approach against any system parameter variations.

VI. CONCLUSION

In this paper, a novel approach is proposed to optimally tune the proportional-integral-derivative (PID) controller gains for load frequency control of renewable integrated hybrid power system using the deep reinforcement learning (DRL) method.

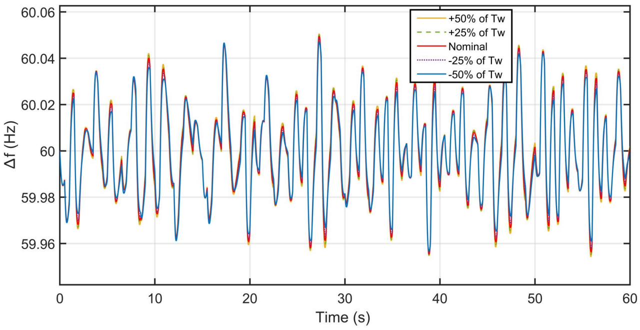

FIGURE 19. Sensitivity analysis response under variation of T_W.

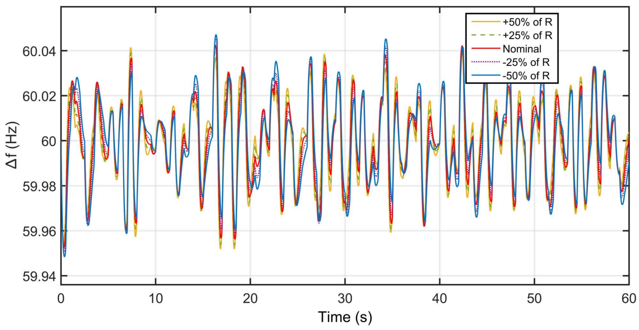

FIGURE 20. Sensitivity analysis response under variation of R.

A twin delayed deep deterministic policy gradient (TD3) algorithm based multi DRL agents are trained, which act as controllers for each area to decide optimal PID values via an interactive trial and error method. The performance of TD3 was compared with deep deterministic policy gradient and meta-heuristic techniques such as genetic algorithm and particle swarm optimization. The results under various scenarios clearly show that our proposed approach outperforms the abovementioned schemes. The TD3 approach gives a significant reduction of almost 50% to 60% in settling time and under/overshoot deviations. All the considered techniques are unable to stabilize the tie-line power under random step load perturbation except the TD3 which certifies its superiority under dynamic variations. Moreover, the proposed scheme greatly compensates the steady-state error and increases the system stability under continuous load-generation variations compared to the conventional control schemes. Furthermore, the sensitivity analysis also indicates that the obtained gain values were robust enough to withstand any system parametric variations.

As the traditional electric grid is undergoing a major transition that incorporates computation in grid operations for better reliability, it brings the key challenge of cyber security. In our future work, we will develop a cyber-attack detection model for LFC to improve the reliability of the system.

APPENDIX

See Figures 21 and 22 here for A1 and A2, respectively.

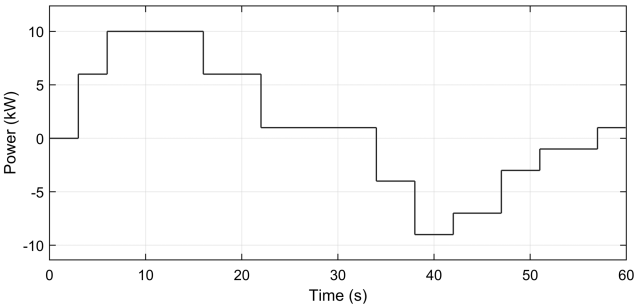

FIGURE 21. Charging/Discharging of electric vehicle.

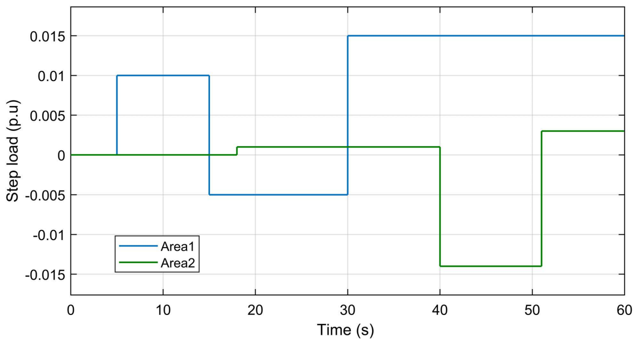

FIGURE 22. Random step load disturbance for both areas.

Rated power = 2000 MW, nominal loading = 1000 MW, (frequency) $f = 60$ Hz, (inertia constant) H = 5, (frequency sensitive load coefficient) D= $\partial P_d/\partial f = 0.00833$, (Area bias parameter) $B_i = 0.425$ p.u.MW/Hz, (Tie-line coefficient) $2\pi T_{12} = 0.545$, (Power system gain) $K_{Pi} = 1/D = 120$, (Power system time constant) $T_{Pi} = 2H/f\ D = 20$s, (Droop constant) $R_i = 2.4$ Hz/p.u.MW, $a_{12} = -1$.

A. HYDRO PLANT

(Hydro governor time constant) $T_h = 48.7$s, $T_{rs} = 0.513$s, $T_{rh} = 10$s, (Penstock water starting time) $T_{pt} = 1$s, $K_H = 0.45$.

B. THERMAL PLANT

(Governor time constant) $T_{gr} = 0.08$s, (Reheater time and gain constant) $T_{re} = 10$s, $K_{re} = 0.33$ Hz/p.u.MW, (steam turbine time constant) $T_{tr} = 0.3$s, $K_T = 0.45$.

C. DIESEL PLANT

$K_D = 16.5$, (Diesel engine speed governing mechanism time constants) T_{d0}, T_{d1}, T_{d2}, $T_{d3} = 1$ s, 2 s, 0.025 s, 3s, $K_D = 0.1$.

D. WIND PLANT

(Time constants of wind turbine) T_{w1}, $T_{w2} = 6$s, 0.041s, (Gain constants of wind turbine) K_{w1}, $K_{w2} = 1.25$, 1.4, Kpc = 0.8, $K_W = 0.25$.

E. PV PLANT

(solar PV time and gain constants) $T_{PV1}, T_{PV2} = 100s, 50s$ and $K_{PV1}, K_{PV1} = -18, 900, K_P = 0.2$.

F. ELECTRIC VEHICLE & BESS

$K_{EV} = 1, T_{EV} = 1s, K_E = 0.1, K_{BESS} = -0.0033, T_{BESS} = 0.1s$.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number IFPRC-190-135-2020 and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

REFERENCES

[1] H. Quan, D. Srinivasan, and A. Khosravi, "Integration of renewable generation uncertainties into stochastic unit commitment considering reserve and risk: A comparative study," *Energy*, vol. 103, pp. 735–745, May 2016.

[2] G. Gross and J. W. Lee, "Analysis of load frequency control performance assessment criteria," *IEEE Trans. Power Syst.*, vol. 16, no. 3, pp. 520–525, Aug. 2001.

[3] H. Bevrani, A. Ghosh, and G. Ledwich, "Renewable energy sources and frequency regulation: Survey and new perspectives," *IET Renew. Power Gener.*, vol. 4, no. 5, pp. 438–457, Sep. 2010.

[4] H. A. Yousef, *Power System Load Frequency Control: Classical and Adaptive Fuzzy Approaches*, 1st ed. Boca Raton, FL, USA: CRC Press, 2017.

[5] S. Aziz, H. Wang, Y. Liu, J. Peng, and H. Jiang, "Variable universe fuzzy logic-based hybrid LFC control with real-time implementation," *IEEE Access*, vol. 7, pp. 25535–25546, 2019.

[6] H. A. Yousef, K. AL-Kharusi, M. H. Albadi, and N. Hosseinzadeh, "Load frequency control of a multi-area power system: An adaptive fuzzy logic approach," *IEEE Trans. Power Syst.*, vol. 29, no. 4, pp. 1822–1830, Jul. 2014.

[7] H. M. Hasanien and S. M. Muyeen, "A Taguchi approach for optimum design of proportional-integral controllers in cascaded control scheme," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1636–1644, May 2013.

[8] S. Kayalvizhi and D. M. V. Kumar, "Load frequency control of an isolated micro grid using fuzzy adaptive model predictive control," *IEEE Access*, vol. 5, pp. 16241–16251, 2017.

[9] J. Guo, "Application of full order sliding mode control based on different areas power system with load frequency control," *ISA Trans.*, vol. 92, pp. 23–24, Sep. 2019.

[10] F. Liu, Y. Li, Y. Cao, J. She, and M. Wu, "A two-layer active disturbance rejection controller design for load frequency control of interconnected power system," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 3320–3321, Jul. 2016.

[11] K. Liao and Y. Xu, "A robust load frequency control scheme for power systems based on second-order sliding mode and extended disturbance observer," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3076–3086, Jul. 2018.

[12] D. C. Das, A. K. Roy, and N. Sinha, "GA based frequency controller for solar thermal–diesel–wind hybrid energy generation/energy storage system," *Int. J. Electr. Power Energy Syst.*, vol. 43, no. 1, pp. 262–279, Dec. 2012.

[13] H. Gozde and M. C. Taplamacioglu, "Automatic generation control application with craziness based particle swarm optimization in a thermal power system," *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 1, pp. 8–16, Jan. 2011.

[14] S. Padhan, R. K. Sahu, and S. Panda, "Application of firefly algorithm for load frequency control of multi-area interconnected power system," *Electric Power Compon. Syst.*, vol. 42, no. 13, pp. 1419–1430, Sep. 2014.

[15] B. P. Sahoo and S. Panda, "Improved grey wolf optimization technique for fuzzy aided PID controller design for power system frequency control," *Sustain. Energy, Grids Netw.*, vol. 16, pp. 278–299, Dec. 2018.

[16] M. Omar, M. Soliman, A. M. A. Ghany, and F. Bendary, "Optimal tuning of PID controllers for hydrothermal load frequency control using ant colony optimization," *Int. J. Electr. Eng. Informat.*, vol. 5, no. 3, pp. 348–360, Sep. 2013.

[17] E. Celik, N. Öztürk, Y. Arya, and C. Ocak, "(1+ PD)-PID cascade controller design for performance betterment of load frequency control in diverse electric power systems," *Neural Comput. Appl.*, vol. 33, no. 22, pp. 15433–15456, 2021.

[18] Y. Arya, N. Kumar, P. Dahiya, G. Sharma, E. Çelik, S. Dhundhara, and M. Sharma, "Cascade-$I^{\lambda}D^{\mu}$N controller design for AGC of thermal and hydro-thermal power systems integrated with renewable energy sources," *IET Renew. Power Gener.*, vol. 15, no. 3, pp. 504–520, Feb. 2021.

[19] M. Glavic, "(Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives," *Annu. Rev. Control*, vol. 48, pp. 22–35, Jan. 2019.

[20] V. P. Singh, N. Kishor, and P. Samuel, "Distributed multi-agent system-based load frequency control for multi-area power system in smart grid," *IEEE Trans. Ind. Electron.*, vol. 64, no. 6, pp. 5151–5160, Jun. 2017.

[21] F. Daneshfar and H. Bevrani, "Load-frequency control: A GA-based multi-agent reinforcement learning," *IET Gener., Transmiss. Distrib.*, vol. 4, no. 1, pp. 13–26, Jan. 2010.

[22] T. Yu, H. Z. Wang, B. Zhou, K. W. Chan, and J. Tang, "Multi-agent correlated equilibrium Q(λ) learning for coordinated smart generation control of interconnected power grids," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 1669–1679, Jul. 2015.

[23] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[24] 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, Mar. 2019.

[25] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process Syst.*, 2017, pp. 6382–6393.

[26] Z. Yan and Y. Xu, "A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system," *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4599–4608, Jun. 2020.

[27] H. Dong and H. Dong, *Deep Reinforcement Learning*. Singapore: Springer, 2020.

[28] J. Li, J. Geng, and T. Yu, "Grid-area coordinated load frequency control strategy using large-scale multi-agent deep reinforcement learning," *Energy Rep.*, vol. 8, pp. 255–274, Nov. 2022.

[29] H. H. Ali, A. M. Kassem, M. Al-Dhaifallah, and A. Fathy, "Multi-verse optimizer for model predictive load frequency control of hybrid multi-interconnected plants comprising renewable energy," *IEEE Access*, vol. 8, pp. 114623–114642, 2020.

[30] C. N. S. Kalyan and G. S. Rao, "Frequency and voltage stabilisation in combined load frequency control and automatic voltage regulation of multiarea system with hybrid generation utilities by AC/DC links," *Int. J. Sustain. Energy*, vol. 39, no. 10, pp. 1009–1029, Nov. 2020.

[31] S. Debbarma and A. Dutta, "Utilizing electric vehicles for LFC in restructured power systems using fractional order controller," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2554–2564, Nov. 2017.

[32] S. Izadkhast, P. Garcia-Gonzalez, and P. Frías, "An aggregate model of plug-in electric vehicles for primary frequency control," *IEEE Trans. Power Syst.*, vol. 30, no. 3, pp. 1475–1482, May 2015.

[33] R. K. Sahu, T. S. Gorripotu, and S. Panda, "A hybrid DE–PS algorithm for load frequency control under deregulated power system with UPFC and RFB," *Ain Shams Eng. J.*, vol. 6, no. 3, pp. 893–911, Sep. 2015.

[34] Y. Arya, "AGC of PV-thermal and hydro-thermal power systems using CES and a new multi-stage FPIDF-(1+PI) controller," *Renew. Energy*, vol. 134, pp. 796–806, Apr. 2019.

[35] S. Kumari and G. Shankar, "Novel application of integral-tilt-derivative controller for performance evaluation of load frequency control of interconnected power system," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 14, pp. 3550–3560, Aug. 2018.

[36] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826–3839, Sep. 2020.

[37] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.

[38] H. Dong, H. Dong, Z. Ding, and S. Zhang, *Deep Reinforcement Learning*. Singapore: Springer, 2020.

JUNAID KHALID received the B.S. degree in electrical engineering from the University of South Asia, Lahore, Pakistan. He is currently pursuing the M.S. degree with the Department of Electrical and Computer Engineering, King Abdulaziz University, Saudi Arabia. He is also a Graduate Research Assistant with the Department of Electrical and Computer Engineering, King Abdulaziz University. His research interests include power systems operation and optimization, sustainable energy systems, and smart grid.

MAKBUL A.M. RAMLI received the B.Eng. degree in electrical engineering from the University of Tanjungpura, Indonesia, in 1995, the M.Eng. degree in electrical engineering from the Bandung Institute of Technology (ITB), Indonesia, in 2000, and the Dr.Eng. degree from the Nagaoka University of Technology (NUT), Japan, in 2005. He is currently a Professor with the Department of Electrical and Computer Engineering, King Abdulaziz University (KAU). His research interests include renewable and alternative energy, distributed generation, energy management systems, and microgrid optimization and control.

MUHAMMAD SAUD KHAN received the B.S. degree in electrical power engineering from IQRA National University, Pakistan, in 2016. He is currently pursuing the M.S. degree in electrical power engineering with the Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include renewable energy and power systems optimization.

TAUFAL HIDAYAT received the B.Eng. and M.Eng. degrees in electrical engineering from the University of Indonesia, in 2012 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, King Abdulaziz University (KAU). He is also a Lecturer with the Institut Teknologi Padang, Indonesia. His research interests include renewable and alternative energy.