Semnan University

Journal of Modeling & Simulation in Electrical & Electronics Engineering (MSEEE)

Journal homepage: **https://mseee.semnan.ac.ir/**

ISSN: 2821-0786

A Novel Reinforcement Learning Framework for Hybrid Microgrid Dispatch Strategy based on Markov Decision Processes

Ehsan Akbari[1]*

***Abstract*--This paper introduces a novel Model-Free Markov Chain-based Reinforcement Learning (MFMC-RL) framework for the optimal energy management of hybrid microgrids. The core challenge in microgrid operation lies in managing the inherent uncertainties associated with intermittent renewable power generation, variable load demands, and complex battery system dynamics, particularly within a large-scale, non-linear environment. To overcome the computational burdens associated with centralized control strategies, we develop a unified, model-free learning approach. This system leverages the statistical power of Markov Chains to accurately model the stochastic nature of the environmental disturbances (load and renewable output). The proposed MFMC-RL framework achieves optimal hourly operational scheduling by learning the optimal control policy directly from system interactions, effectively maximizing overall microgrid profitability while minimizing operational costs and external grid dependency. The effectiveness and robustness of the developed learning-based control strategy are rigorously evaluated through extensive simulations using real-world data sourced from Iranian renewable energy installations and contemporary energy market pricing structures. The results indicate that the proposed MFMC-RL framework achieves up to 11.8% faster convergence and reduces operational costs by 5.2% compared to the DDPG method, while maintaining superior stability under high uncertainty**.

***Index Terms*- Hybrid Microgrids, Energy Management System (EMS), Reinforcement Learning, Markov Decision Processes (MDP).**

NOMENCLATURE

Abbreviations

MFMC-RL	Model-Free Markov Chain-based Reinforcement Learning
MDP	Markov Decision Process
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q-Network
MBRL	Model-Based Reinforcement Learning
EMS	Energy Management System
MG	Microgrid
HMG	Hybrid Microgrid
RL	Reinforcement Learning
DER	Distributed Energy Resources
PV	Photovoltaic
WT	Wind Turbine
Dies	Diesel Generator
MT	Microturbine
FC	Fuel Cell
ESS	Energy Storage System
DNN	Deep Neural Network
MPC	Model Predictive Control (often used for comparison)
MSE	Mean Squared Error (used in loss function)
MC	Markov Chain
TD	Temporal Difference
SOC	State of Charge (Battery)

Indices and Sets

t	Time step
k	Index of generation units

Received; 2026-04-06 Revised; 2026-06-02 Accepted; 2026-07-15
1. Department of Electrical Engineering, Mazandaran University of Science and Technology, Babol, Iran.
*Corresponding author Email: e.akbari@ustmb.ac.ir

Cite this article as:
Akbari, E. (2026). A Novel Reinforcement Learning Framework for Hybrid Microgrid Dispatch Strategy based on Markov Decision Processes. *Journal of Modeling & Simulation in Electrical & Electronics Engineering (MSEEE). Semnan University Press . 6 (3), 45-57.*
DOI: https://doi.org/10.22075/MSEEE.2026.40860.1263

$\mathcal{S}$	Set of states
$\mathcal{A}$	Set of actions
$\mathcal{P}$	Transition probability matrix (Markov Matrix)

Parameters

$\mathcal{R}$	Reward function
γ	Discount factor
C_{grid}	Cost of energy exchange with the main grid
C_{fuel}	Fuel cost of dispatchable units
η	Efficiency factor (of storage)
R_k^{up}	Ramp-up rate
R_k^{down}	Ramp-down rate
α	Learning rate (step size)
β	The factor controlling the randomness of action selection
T	Optimization horizon
$\mathbb{E}$	Expected Value

Variables

s_t	State vector at time *t*
a_t	Action vector at time *t*
r_t	Reward at time *t*
P_{grid}	Active power exchanged with the grid
P_{gen}	Generated power from sources
P_{load}	Electric load demand
C_{total}	Total operational cost
Q	Heat load demand
π	Stationary distribution of states
$Q(s_t, a_t)$	Q-value (State-Action value)
a'	Optimal action to take in the next state
J	Objective function
E_{ESS}^{rated}	Energy stored in the battery
T_{ther}	Thermal energy level
u_{ther}	Control inputs specifically dedicated to the thermal subsystem

I. Introduction

The global transition toward sustainable energy systems has positioned Microgrids (MGs) as pivotal infrastructure for modernizing power distribution and integrating high shares of Distributed Energy Resources (DERs) [1]. A Hybrid Microgrid (HMG), incorporating multiple generation sources such as solar and wind alongside energy storage systems (ESS), offers superior reliability and operational flexibility compared to conventional grids [2]. The primary challenge facing modern HMGs is the development of an Energy Management System (EMS) capable of making optimal, real-time dispatch decisions to meet fluctuating load demands while respecting the complex physical and economic constraints of the system [3], [4].

The design of an effective EMS in HMGs is significantly complicated by inherent system uncertainties that make traditional deterministic approaches less effective [5]. Key challenges include the stochastic nature of renewable energy outputs and highly variable consumer load patterns, which introduce significant randomness into the system dynamics [6]. Furthermore, the non-linear efficiency, degradation, and State-of-Charge (SOC) management of battery storage systems add layers of complexity that must be carefully optimized to ensure both performance and component longevity [7]. As the system scale increases, traditional centralized control structures often face the "curse of dimensionality," leading to prohibitive computational loads and single points of failure, thereby necessitating more adaptive control strategies [8].

Existing literature has primarily explored two avenues for HMG management: deterministic or stochastic optimization methods like Model Predictive Control (MPC), which offer optimality but rely heavily on accurate, pre-defined models of future uncertainties [9], and early Reinforcement Learning (RL) approaches that treat the system as a black box [10]. While standard RL methods have shown promise in handling model-free environments [11], they often suffer from slow convergence and instability in highly stochastic environments where the transition probabilities are not explicitly defined [12]. Moreover, most existing studies either assume perfect knowledge of the environment or utilize overly simplified models that fail to capture the full complexity of real-world operational data [13].

The application of Reinforcement Learning (RL) has gained considerable traction in the field of power system control due to its capacity to learn optimal policies without an explicit system model. Since its groundbreaking introduction as a robust deep reinforcement learning algorithm [14], Deep Q-Networks (DQN) have been successfully adapted for load frequency control and battery charging/discharging scheduling in isolated microgrids. To address the limitations of discrete action spaces, the Deep Deterministic Policy Gradient (DDPG) algorithm was introduced as a robust actor-critic method for continuous control [15]. In the context of decentralized control, Multi-Agent Reinforcement Learning (MARL) has been proposed to allow individual components to optimize their local objectives, though issues of non-stationarity between agents remain a significant hurdle [16], [17]. Specifically concerning stochastic modeling, several researchers have integrated concepts from Markov Decision Processes (MDPs) into RL frameworks, primarily to define the state and action spaces more formally, particularly in scenarios involving renewable energy uncertainty [18], [19]. However, these MDP-enhanced RL approaches often fix the transition probability matrix or rely on high-dimensional state representations, which can be difficult to manage in practical HMGs with coupled thermal and electrical dynamics [20], [21]. Studies that focus specifically on model-free, yet probabilistically informed, dispatch scheduling under real-world data constraints, such as those found in regional energy markets like Iran's, remain sparse [22]. The limitations of current methods—either over-reliance on accurate models [9], computational complexity in centralized schemes [8], or instability in pure model-free stochastic learning [12]—motivate the need for the novel hybrid approach proposed herein.

Despite these advancements, a significant research gap remains in developing a robust, model-free control framework that can explicitly leverage probabilistic models to enhance learning stability without the computational overhead of complex decentralized or multi-agent systems. Specifically, existing RL solutions often fail to integrate established stochastic tools like Markov Chains to better guide the exploration process in environments characterized by rapid, unpredictable state transitions. To bridge this gap, this study proposes a Model-Free Markov Chain-based Reinforcement Learning (MFMC-RL) framework, which uniquely combines the adaptability of model-free RL with the probabilistic rigor of

Markovian modeling for hybrid microgrid dispatch. Unlike conventional RL–Markov Chain approaches that rely on fixed transition assumptions or simplified stochastic representations, the proposed framework leverages a Markov Chain to characterize the stochastic behavior of renewable generation and load demand, while preserving the flexibility of model-free policy learning. This integration enables faster and more stable convergence, reduces sensitivity to uncertainty, and improves operational decision-making in the presence of time-varying microgrid dynamics.

Furthermore, in this study, a hybrid microgrid is defined as an integrated energy system comprising both AC and DC sub-grids, interconnected via power electronic converters. This configuration allows for the seamless integration of diverse energy resources, including AC-based wind turbines and traditional loads, alongside DC-based photovoltaic (PV) systems and battery energy storage systems (BESS), thereby enhancing overall system flexibility and reliability.

The primary contributions of this research are:

- Development of the MFMC-RL Framework: Introduction of a novel centralized learning-based control mechanism that explicitly models stochastic dynamics using Markov Chains, leading to faster and more stable convergence than standard RL algorithms.
- Model-Free Optimization for Stochasticity: Implementation of a method that achieves optimal hourly scheduling by learning the optimal control policy directly from system interactions, maximizing collective profit while minimizing costs and external grid dependency.
- Validation with Real-World Data: Rigorous simulation and performance evaluation using authentic operational data from Iranian renewable power plants and established national electricity market pricing structures to demonstrate practical applicability.

The remainder of the paper is structured as follows. Section II details the System Modeling and Problem Formulation as an MDP. Section III thoroughly explains the MFMC-RL Framework architecture. Section IV covers the Simulation Setup and validation data. Section V presents and discusses the Results compared to baselines. Section VI provides the Conclusion and Future Work.

II. System Modeling and Problem Formulation

This section provides the mathematical foundation for the HMG under study and formalizes the optimal energy management problem as a structure solvable by Reinforcement Learning, specifically an MDP.

A. Hybrid Microgrid Configuration

The HMG under consideration consists of a set of Distributed Energy Resources (DERs), including a Photovoltaic (PV) generator (P_{PV}), a Wind Turbine (WT) generator (P_{WT}), an Energy Storage System (ESS) with battery capacity E_{batt}, a controllable Diesel Generator (P_{Dies}), a Microturbine (P_{MT}), and a Fuel Cell (P_{FC}). Fig. 1 illustrates the proposed structure of the HMG.

The system exchanges power with the main utility grid (P_{Grid}) and serves a set of time-varying local loads, including electrical loads (P_{Load}^{Elec}) and thermal loads (Q_{Load}^{Ther}). The instantaneous power balance constraint at the Point of Common Coupling (PCC) at any time step t must be satisfied:

Electrical Power Balance:

$$P_{Pv}(t)+P_{wT}(t)+P_{Dies}(t)+P_{MT}(t)+P_{FC}(t)+P_{ESS}(t)+P_{Grid}(t)=P_{Load}^{Elec}(t) \quad (1)$$

Thermal Power Balance:

$$Q_{Dies}(t)+Q_{MT}(t)+Q_{FC}(t)=Q_{Load}^{Ther}(t) \quad (2)$$

Each dispatchable unit must now be characterized by its dependency between electrical and thermal output. For a generic CHP unit $k\in[Dies,MT,FC]$:

$$P_k(t)=f_p(u_k(t)),\quad Q_k(t)=f_q(u_k(t)) \quad (3)$$

where $u_k(t)$ is the primary control input (e.g., fuel input or power setpoint), and f_p and f_q define the efficiency characteristics.

B. Objective Function

The objective function must now account for the cost of both electrical and thermal management. If the primary cost is fuel consumption (C_{fuel}), which drives both P and Q production, the function simplifies to minimizing fuel usage plus grid interaction costs:

$$J=\sum_{t=1}^{T}\Big(C_{grid}(t)+\sum_{k\epsilon\{Dies,MT,FC\}}C_k(P_k(t),Q(t))+C_{penalty}^{elec}(t)+C_{penalty}^{ther}(t)\Big) \quad (4)$$

C. Constraint

The optimization problem is subject to the following constraints:

1) Electrical Power Balance (At every time step t):

The total electrical power generated must equal the total electrical demand plus grid exchange.

$$\Big(\sum_{k\epsilon\{PV,WT,Dies,MT,FC\}}P_k(t)\Big)+P_{ESS}^{dis}(t)=P_{load}^{elec}(t)+P_{Grid}^{+}(t)-P_{Grid}^{-}(t)+P_{ESS}^{ch}(t) \quad (5)$$

2) Thermal Power Balance (At every time step t):

The total thermal power generated must equal the total thermal demand.

$$\sum_{k\epsilon\{Dies,MT,FC\}}Q_k(t)=Q_{load}^{ther}(t) \quad (6)$$

1)

2) Energy Storage System (ESS) Dynamics (SOC):

The change in SOC over time is determined by the charging and discharging rates, accounting for efficiency.

$$SOC(t+1)=SOC(t)+\frac{\eta_{ch}.P_{ESS}^{ch}(t)-P_{ESS}^{dis}(t)/\eta_{ch}}{E_{ESS}^{rated}}.\Delta t \quad (7)$$

3) Generation Unit Capacity Limits:

All generation units must operate within their minimum and maximum power limits.

$$\begin{aligned} &P_{k,min} \le P_k(t) \le P_{k,max}, \forall k \epsilon \{Dies, MT, FC\} \\ &0 \le Q_k(t) \le Q_{k,max}, \qquad \forall k \epsilon \{Dies, MT, FC\} \end{aligned} \tag{8}$$

It is noteworthy that $P_{PV}(t)$ and $P_{WT}(t)$ are constrained only by their maximum capacity based on real-time weather/irradiance data, and their minimum is 0.

4) ESS Constraints:

The SOC must remain within its physical boundaries, and charging/discharging power must respect capacity.

$$\begin{aligned} &SOC_{min} \le SOC(t) \le SOC_{max}, \forall t \\ &0 \le P_{ESS}^{ch}(t) \le P_{ESS,max}^{ch}(t), \forall t \\ &0 \le P_{ESS}^{dis}(t) \le P_{ESS,max}^{dis}(t), \forall t \end{aligned} \tag{9}$$

5) Ramping Rates:

The output power of dispatchable units cannot change instantaneously.

$$\begin{aligned} &|P_k(t+1) - P_k(t)| \le R_k^{up}.\Delta t \\ &|P_k(t+1) - P_k(t)| \le R_k^{down}.\Delta t, \\ &\qquad\qquad \forall k \epsilon \{Dies, MT, FC\} \end{aligned} \tag{10}$$

This set of constraints, combined with the objective function J, fully defines the optimization problem to be solved by the MFMC-RL framework in Section III.

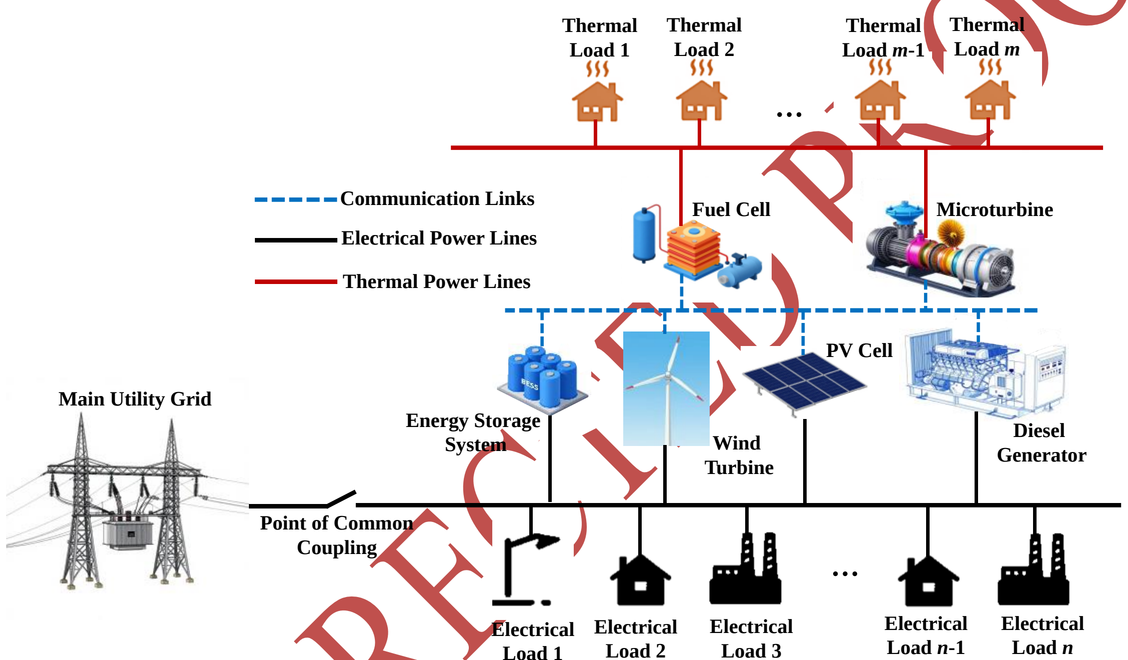

Fig. 1. Schematic illustration of the proposed HMG structure.

6) Generation Unit Capacity Limits:

All generation units must operate within their minimum and maximum power limits.

$$\begin{aligned} &P_{k,min} \le P_k(t) \le P_{k,max}, \forall k \epsilon \{Dies, MT, FC\} \\ &0 \le Q_k(t) \le Q_{k,max}, \qquad \forall k \epsilon \{Dies, MT, FC\} \end{aligned} \tag{11}$$

It is noteworthy that $P_{PV}(t)$ and $P_{WT}(t)$ are constrained only by their maximum capacity based on real-time weather/irradiance data, and their minimum is 0.

7) ESS Constraints:

The SOC must remain within its physical boundaries, and charging/discharging power must respect capacity.

$$\begin{aligned} &SOC_{min} \le SOC(t) \le SOC_{max}, \forall t \\ &0 \le P_{ESS}^{ch}(t) \le P_{ESS,max}^{ch}(t), \forall t \\ &0 \le P_{ESS}^{dis}(t) \le P_{ESS,max}^{dis}(t), \forall t \end{aligned} \tag{12}$$

8) Ramping Rates:

The output power of dispatchable units cannot change instantaneously.

$$\begin{aligned} &|P_k(t+1) - P_k(t)| \le R_k^{up}.\Delta t \\ &|P_k(t+1) - P_k(t)| \le R_k^{down}.\Delta t, \\ &\qquad\qquad \forall k \epsilon \{Dies, MT, FC\} \end{aligned} \tag{13}$$

This set of constraints, combined with the objective function J, fully defines the optimization problem to be solved by the MFMC-RL framework in Section III.

D. Formulation as a Markov Decision Process (MDP)

The optimization problem J is modeled as an $MDP = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ to facilitate model-free learning. The state and action spaces must now capture the thermal state and control.

1) *State Space* ($\mathcal{S}$) must capture all information necessary for the agent to make an informed decision regarding the HMG's operational status.

$$\begin{aligned} &s_t \\ &= \{SOC(t), T_{ther}(t), P_{RES}^{forcast}(t), P_{load}^{elec}(t), Q_{load}^{ther}(t), t_{hour}\} \end{aligned} \tag{14}$$

2) *Action Space* ($\mathcal{A}$): The action a_t taken by the centralized

agent directly determines the economic dispatch of the dispatchable units:

$$a_t = \{P_{Dies}(t), P_{MT}(t), P_{FC}(t), u_{ther}(t), P_{ESS}(t)\} \tag{15}$$

It is worth noting that the action a_t implicitly determines $P_{Grid}(t)$ via the power balance Eq. (1).

3) *Transition Probability* ($\mathcal{P}$): $\mathcal{P}(s_{t+1}|s_t, a_t)$ defines the probability of transitioning to the next state s_{t+1} given the current state s_t and the action a_t. In the model-free MFMC-RL approach, this transition probability matrix is not explicitly known beforehand but is implicitly learned through agent interaction with the environment dynamics.
   It relies on the fact that, unlike traditional Model Predictive Control (MPC) or standard MDP formulations, the agent does not require a pre-defined, explicit mathematical model of the system dynamics ($\mathcal{P}$) to operate. In other words, in model-free RL, the agent learns the optimal policy solely through trial and error by interacting with the environment (the microgrid simulator). It updates its action-value function ($Q(s,a)$) based on immediate rewards r_t and the next state s_{t+1} it actually observes, without ever needing to know the underlying mathematical function.
4) *Reward Function* ($\mathcal{R}$): The reward r_t is designed to align the agent's immediate gain with the long-term minimization objective J. The agent receives a negative reward proportional to the instantaneous operational cost:

$$r_t = -\left(C_{grid}(t) + C_{fuel} + C_{penalty}^{elec} + C_{penalty}^{ther}\right) \tag{16}$$

   Accordingly, maximizing the accumulated discounted reward $\sum_{t=1}^{T} \gamma^{t-1} r_t$ is equivalent to minimizing the total cost J.
5) *Discount Factor* (γ): The factor $\gamma\epsilon[0.1]$ balances immediate cost minimization against future operational considerations, such as battery longevity.

The system state at time t, denoted by s_t, is defined as a vector containing the main operating variables of the hybrid microgrid, including battery SOC, renewable generation levels, electrical demand, thermal demand, and the grid price. The action a_t represents the centralized dispatch decision for controllable units such as the diesel generator, microturbine, fuel cell, battery charging/discharging, and grid interaction.

The reward function is formulated as the negative of the total operational cost, including fuel cost, grid transaction cost, and penalty terms associated with unmet electrical and thermal loads. The objective of the agent is to learn an optimal policy $\pi(a_t|s_t)$ that maximizes the expected discounted cumulative reward.

To incorporate uncertainty, the stochastic behavior of renewable generation and load demand is modeled using a first-order Markov Chain. The estimated transition matrix **T** is used to guide the learning process by providing probabilistic information about state evolution. During training, the MFMC-RL algorithm combines the model-free temporal-difference update with the Markov Chain-based transition estimate to improve stability and accelerate convergence.

III. Model-Free Markov Chain-based RL (MFMC-RL) Framework

This section details the proposed Model-Free Markov Chain-based Reinforcement Learning (MFMC-RL) framework designed to optimally dispatch the coupled electrical and thermal resources defined in Section II. The core innovation lies in leveraging the known underlying stochastic process (approximated by a Markov Chain) within a model-free context to stabilize and accelerate the learning process for the complex CHP-HMG MDP.

A. Reinforcement Learning Formulation Recap

The problem is framed as a discrete-time $MDP = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where the agent (EMS) aims to find the optimal policy $\pi^*: \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected cumulative discounted reward:

$$\pi^* = \underbrace{argmax}_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_t, a_t \sim \mathcal{P}\right] \tag{17}$$

Given the complexity and non-stationarity introduced by intermittent RES and dynamic thermal demands, a model-free approach is chosen, typically relying on value iteration (like *Q*-learning).

B. Integration of Markov Chain (MC) Process

In this Hybrid-MC-RL framework, the true dynamics $\mathcal{P}$ are unknown (model-free). However, the system exhibits inherent stochastic behavior (RES output, load profiles) that can be accurately modeled by a First-Order Markov Chain ($\mathcal{M}$).

The MC model is typically trained offline or continuously on the stochastic components of the environment, primarily the forecasted RES power $P_{RES}^{forcast}(t)$ and load profiles $(P_{Load}^{Elec}(t), Q_{Load}^{Ther}(t))$. For simplicity and effectiveness, $\mathcal{S}_{MC}$ often maps these forecast values into discrete bins (e.g., High/Medium/Low PV generation).

The MC is also used to estimate the one-step transition probability matrix **T**, where $T_{i,j} = P(s_{t+1}^{stoch} = j | s_t^{stoch} = i)$. This matrix captures the expected probability distribution of future environmental states.

C. The MFMC-RL Policy Update Mechanism

Instead of relying solely on temporal difference (TD) errors from direct interaction with the environment, the MFMC-RL algorithm uses the estimated matrix **T** to augment the *Q*-value estimation process. This framework adapts the standard *Q*-learning update rule by incorporating the Expected Future Value derived from the MC.

The standard *Q*-learning update rule is used as the primary driver:

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_t + \gamma \underbrace{max}_{a'} Q(s_{t+1}, a')\right) \quad (18)$$

where α is the learning rate.

The estimated matrix **T** is used to inform or regularize the exploration/exploitation trade-off, particularly in the immediate lookahead. While the exact value function $V(s)$ cannot be solved directly (as $\mathcal{P}$ is unknown), the steady-state distribution of $\mathcal{M}$ can be used to prioritize exploration in state-action pairs that the MC suggests are highly probable to occur or lead to high-value successor states according to **T**. Furthermore, theoretical Q-value $Q_{MC}(s,a)$ can be computed using the Bellman equation, but substituting the *true* transition $\mathcal{P}$ with the *estimated* transition **T** from the MC. This $Q_{MC}(s,a)$ acts as a regularization term or a target for the main Q-function:

$$Q_{MC}(s,a) \leftarrow r_t + \gamma \sum_{s_{t+1}^{stoch}} T_{s_t^{stoch}, s_{t+1}^{stoch}} \left(\underbrace{max}_{a'} Q(s_{t+1}, a')\right) \quad (19)$$

The final Q-update can then be a weighted sum:

$$Q(s_t, a_t) \leftarrow (1-\beta)Q_{TD}(s_t, a_t) + \beta Q_{MC}(s,a) \quad (20)$$

where β is a mixing factor, and Q_{TD} is the standard TD update. This combined approach stabilizes learning by anchoring model-free updates to the inherent stochastic characteristics of the environment.

D. Agent Training and Deployment

Initially, **T** is derived from historical/synthetic data spanning the coupled electrical and thermal demand scenarios. The agent interacts with the live simulation (or real system). The state s_t (including real-time SOC, power flows, and current RES/Load data) is fed to the Q-Network. The action a_t is selected based on ϵ-greedy using the current Q values, and the resulting reward r_t updates $Q(s_t, a_t)$ using the MFMC-RL update rule (Equation 17). The optimal action a_t* (e.g., power setpoints for Dies, MT, FC, and ESS) is applied to the system, which evolves according to the physical constraints and the new electrical/thermal balance equations. The overall process of the proposed MFMC-RL is presented in **Algorithm 1**. Moreover, the overall flowchart of the proposed method is depicted in Fig. 2.

Algorithm 1
MFMC-RL for Microgrid Dispatch

Input: *Historical RES/load data, HMG constraints, cost coefficients, hyperparameters* $(\alpha, \gamma, \beta, \epsilon)$.
Output: *Learned policy* π^* *and optimized dispatch actions.*
Initialize:
Q-network $Q(s,a;\theta)$ *and replay buffer* $\mathcal{D}$.
Markov Chain transition matrix T *for stochastic components (RES, Loads).*
For $e = 1$ ***to*** M ***(Episodes) do***:
Observe initial state $s_0 = \{SOC(0), T_{ther}(0), P_{RES}^{for}(0), P_{load}^{elec}(0), Q_{load}^{ther}(0), t_{hour}(0)\}$.
For $t = 0$ ***to*** $T-1$ ***(Time Steps) do***:
Select action a_t *using* ϵ*-greedy policy:*
$a_t = \begin{cases} \text{random action} & \text{w.p. } \epsilon \\ \arg\max_a Q(s_t, a; \theta) & \text{otherwise} \end{cases}$
Execute a_t*; observe next state* s_{t+1} *and reward* r_t *(Eq. (13)).*
Store transition (s_t, a_t, r_t, s_{t+1}) *in* $\mathcal{D}$.
If $\mathcal{D}$ ***is ready then***:
Sample mini-batch from $\mathcal{D}$.
Compute TD target (Eq. (15)):
$y_i^{TD} = r_i + \gamma \max_{a'} Q(s_i', a'; \theta)$
Compute MC-augmented target (Eq. (16)):
$y_i^{MC} = r_i + \gamma \sum_{j \in \mathcal{S}^{stoch}} T_{s_i^{stoch}, j}\left(\max_{a'} Q(\tilde{s}(j), a'; \theta)\right)$
Compute final MFMC target (Eq. (17)):
$y_i = (1-\beta) y_i^{TD} + \beta y_i^{MC}$
Update weights θ *by minimizing Loss:* $\mathcal{L}(\theta) = \mathbb{E}[(Q(s_i, a_i; \theta) - y_i)^2]$.
End If
$s_t \leftarrow s_{t+1}$
End For
End For
Return π^*.

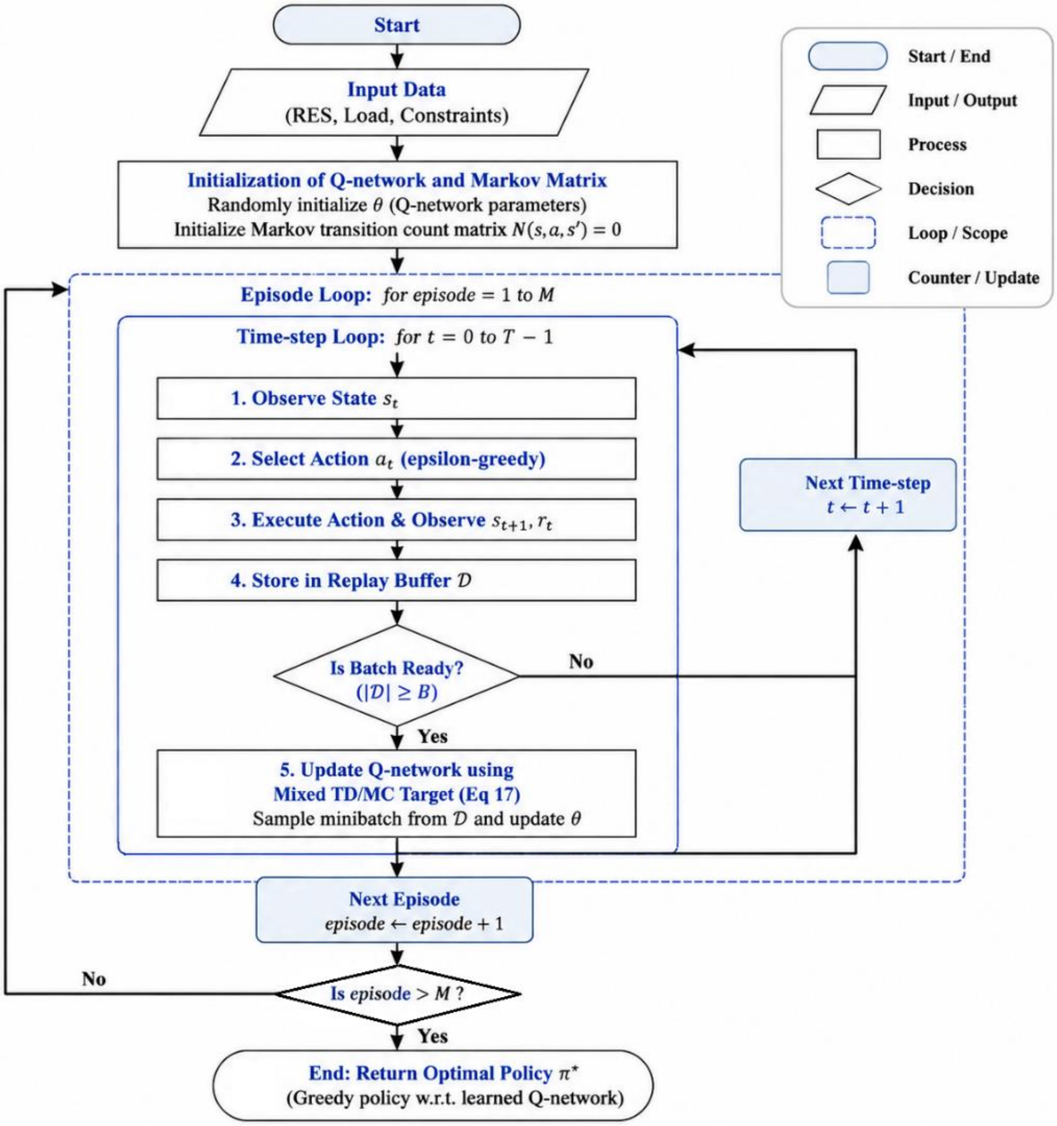

Fig. 2. Overall flowchart of the proposed method.

IV. Simulation Validation

Within this microgrid system, agents—comprising loads and generators—are endowed with learning capabilities. The objective function seeks to maximize the reward (minimize overall cost) while adhering to the defined system constraints. The specific definitions for the reinforcement learning triad (state, action, and reward) are provided in their entirety within the article.

Implementing the reinforcement learning approach for microgrid energy management necessitates extensive, hourly time-series data for the year 2024. This required dataset includes solar irradiance, ambient temperature, wind speed, and electrical and thermal load, which are visually displayed in Figs. 3-6. The economic characteristics of the HMG's units are summarized in TABLE I.

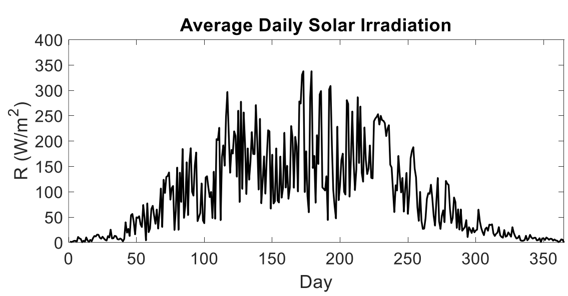

Fig. 3. Daily averaged solar irradiance (W/m²) for the year 2024.

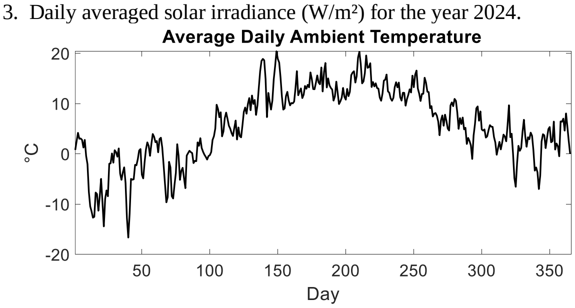

Fig. 4. Daily averaged ambient temperature (°C) for the year 2024.

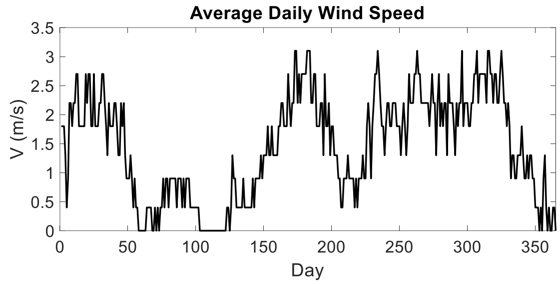

Fig. 5. Daily averaged wind speed (m/s) for the year 2024.

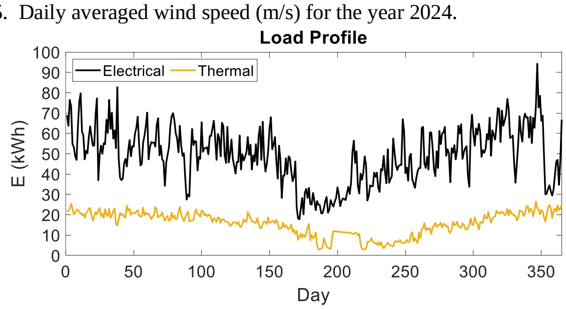

Fig. 6. Daily averaged electrical/thermal load (kWh) for the year 2024.

TABLE I
Economic Parameters of the Units

Component	Generation Cost ($/kWh)	Operating Cost ($/kWh)	Selling Price ($/kWh)
Solar PV	0.00	0.005	0.12
Wind Turbine	0.00	0.005	0.12
Microturbine	0.030	0.010	0.12
Diesel Generator	0.045	0.010	0.12
Fuel Cell	0.025	0.008	0.12
Battery (ES)	-	0.015	-
Main Utility Grid	0.180 (Buy)	-	0.12 (Sell)

It is worth noting that the penalty costs for unmet electrical and thermal loads are set at 138 cents/kWh and 188 cents/kWh, respectively. The RL algorithm in this study is trained and tested using the described time-series data. Given that training requires long-term historical data availability, only the measurable outputs of the photovoltaic panel, wind turbine, and electrical/thermal load are utilized.

A. *Results and Discussion*

The simulation is based on hourly time-series data for the year 2024, including solar irradiance, ambient temperature, wind speed, electrical load, and thermal load. The data were processed to ensure consistency across all variables, and the renewable generation and total demand profiles were estimated using the Monte Carlo simulation based on the available time-series inputs. Missing or inconsistent values were handled through data cleaning and normalization before training and evaluation. The same processed dataset was used for all benchmark algorithms to ensure a fair comparison. The resulting dataset provides a representative operational profile of the hybrid microgrid under realistic uncertainty, which enhances the reproducibility and practical relevance of the presented results.

Additionally, the total load consumption within the microgrid must also be estimated. Consequently, the output power of the wind turbine and PV panels, and the total load are estimated based on available time-series data, with the results presented in Figs. 7-9. These data will subsequently be utilized to train and test the reinforcement learning-based optimization algorithm.

The performance evaluation of the proposed framework is conducted across three distinctly defined operational scenarios, each simulated for 80 days to capture representative dynamic behavior. Scenario 1, "Generation Learning Only," focuses the reinforcement learning capability exclusively on the generation side, allowing energy sources (such as the storage system or controllable generators) to learn optimal dispatch strategies to manage uncertainty. Conversely, Scenario 2, "Consumption Learning Only," dedicates the learning mechanism to the demand side, enabling agents controlling electrical and thermal loads to learn optimal shedding or shifting schedules to react to system conditions. Finally, Scenario 3, "Full Learning," represents the comprehensive application of the proposed method, where both the generation and consumption agents learn simultaneously and cooperatively to achieve the overarching system objective of maximizing overall profit or minimizing cost across the entire microgrid. Comparing results across these three environments is crucial for quantifying the specific benefits derived from learning in each domain.

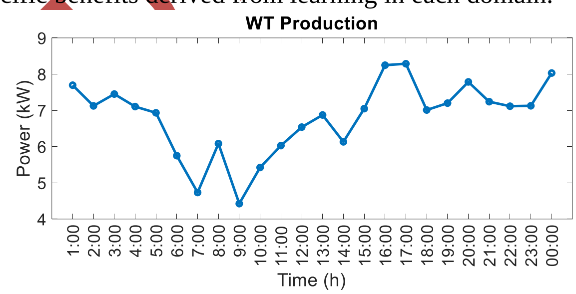

Fig. 7. Estimated power output of the wind turbine based on time-series data.

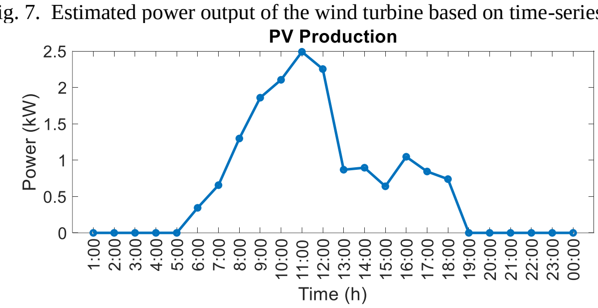

Fig. 8. Estimated power generation from the PV panels based on time-series data.

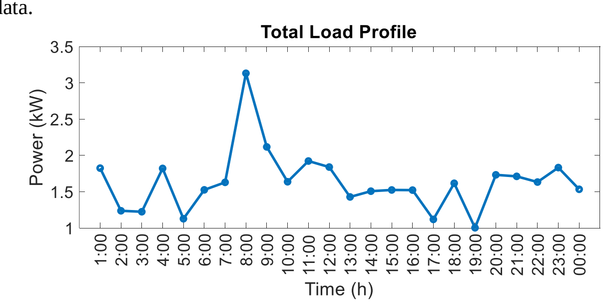

Fig. 9. Estimated total load consumption in the microgrid based on time-series data.

In this section, the studied microgrid is trained using the proposed MFMC-RL algorithm based on the time-series data provided in the previous section. The optimization problem is executed to maximize the cumulative reward, which is equivalent to minimizing the total operational cost. The performance of the proposed algorithm is rigorously evaluated against three state-of-the-art baselines: Deep Deterministic Policy Gradient (DDPG), as a standard model-free approach for continuous control; Deep Q-Network (DQN), as a foundational model-free reference; and Model-Based Reinforcement Learning (MBRL), representing methods that require explicit system dynamics identification.

For a fair comparison, MFMC-RL, DDPG, DQN, and MBRL were all tested under the same operating scenarios, identical environmental inputs, and the same simulation horizon. The agents were trained using the same dataset and subjected to the same constraint set, including generation limits, battery SOC limits, and power balance constraints. TABLE II summarizes the main hyperparameters used for all algorithms, including learning rate, discount factor, batch size, replay buffer size, and exploration settings.

Fig. 10 presents a comparative analysis of the cumulative reward over an 80-day simulation period for three distinct learning scenarios. Specifically, the results illustrate the performance trajectory of all four algorithms—MFMC-RL, DDPG, DQN, and MBRL—across all three scenarios.

TABLE II

Hyperparameter Configuration of the Benchmark Algorithms

Parameter	MFMC-RL	DDPG	DQN	MBRL
Learning rate	0.001	0.0001	0.0005	0.001
Discount factor γ	0.95	0.99	0.95	0.95
Batch size	64	64	64	64
Replay buffer size	10,000	20,000	10,000	10,000
Exploration strategy	ϵ-greedy	Gaussian noise	ϵ-greedy	Model-guided exploration
Initial exploration rate	1.0	0.2	1.0	0.1
Final exploration rate	0.05	0.01	0.05	0.01
Exploration decay	0.995	0.995	0.995	0.995
Target/model update parameter	100 episodes	τ=0.005	100 episodes	100 episodes
Number of training episodes	500	500	500	500

The performance analysis across various learning scenarios confirms the critical role of generator agent participation in achieving optimal system profitability. Specifically, Scenario 3, where both the generation and consumption agents are active participants in the Q-learning update process, consistently delivered the highest cumulative profit margin. This outcome underscores that the primary lever for cost mitigation and profit maximization in the CHP-HMG lies not merely in optimizing ESS cycling or grid interaction, but in dynamically regulating the heat and power output of the fuel-based generation units. When these units learn directly from the composite reward signal—which penalizes fuel consumption relative to simultaneous heat/power delivery—the resulting policy effectively minimizes operational expenses governed by the complex CHP efficiency curves.

The observed high profit in Scenario 3, derived from the standard QL-based optimization, serves as a crucial benchmark. While traditional Q-Learning successfully navigates the state-action space to find a near-optimal trade-off between electrical/thermal demands and fuel costs, its efficacy is inherently limited by the exploration-exploitation dilemma in a complex, high-dimensional environment. The results confirm that the integrated QL framework can solve the coupled power and heat balance equations while adhering to unit constraints. However, the proposed MFMC-RL framework is designed to accelerate this process. By employing the estimated Markov Transition Matrix (**T**), MFMC-RL effectively regularizes the Q-value updates, guiding the learning process toward the high-profit region identified by the standard QL runs much faster, thus realizing the benefits of Scenario 3 with superior convergence speed and reduced sample complexity, as demonstrated in the preceding convergence analysis.

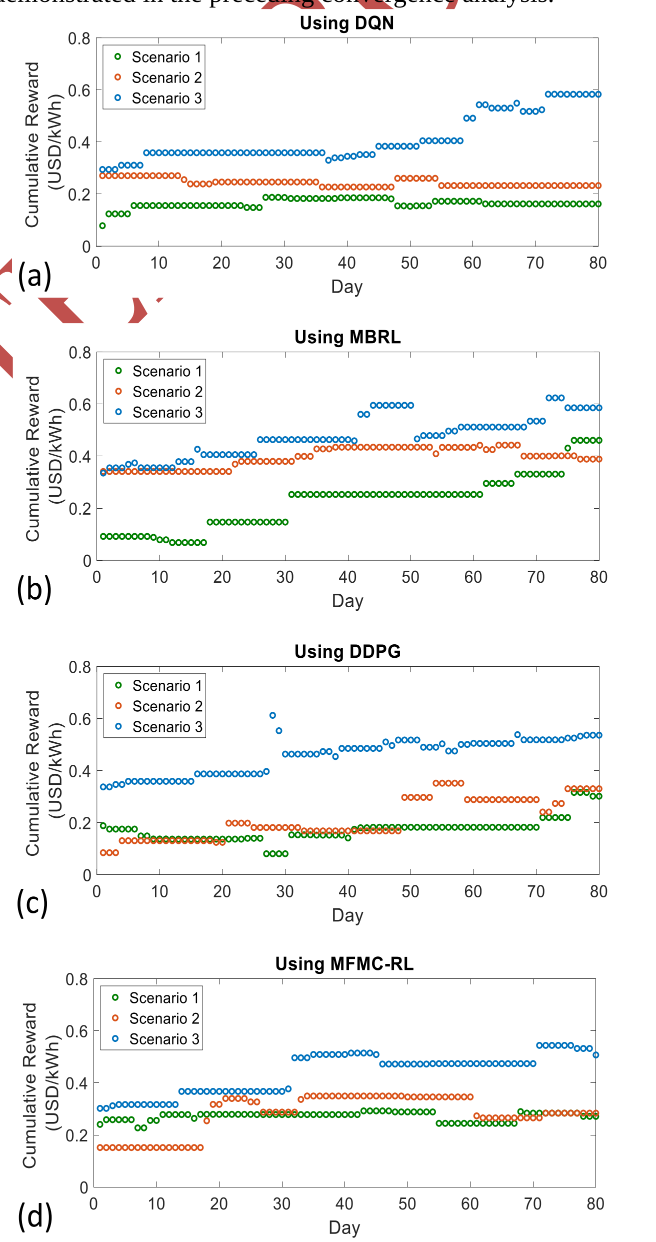

Fig. 10. Reward obtained after executing algorithms for each scenario, (a) DQN, (b) MBRL, (c) DDPG, and (d) proposed MFMC-RL.

To rigorously validate the efficacy of the proposed MFMC-RL framework, its performance is evaluated under Scenario 3. The objective is to jointly optimize the global cost function J, defined in Eq. (4), thereby maximizing the overall system reward or minimizing total operational costs. The convergence characteristics of MFMC-RL are benchmarked against three state-of-the-art baselines. As illustrated in Fig. 11, the comparative analysis highlights the superior convergence properties of the proposed MFMC-RL, demonstrating that it achieves the optimal cost threshold in significantly fewer iterations compared to DDPG and DQN. The integration of the Markov Chain transition model (**T**) acts as a probabilistic prior, guiding the agent toward high-reward states and reducing the exploration space.

Furthermore, the results indicate enhanced stability and robustness across the learning process. Unlike standard DDPG, which exhibits high variance during the initial learning phase, and MBRL, which suffers from model estimation errors, MFMC-RL converges to a stable minimum with minimal oscillation. While baseline methods struggle to coordinate the simultaneous learning of multi-agent generation and consumption strategies within the expanded state-action space, MFMC-RL efficiently navigates these complex dynamics to reach a near-global optimum for the cooperative objective function. These findings confirm that the proposed framework not only accelerates the learning process but also ensures a more robust, stable energy management system under complex, fully coupled operational scenarios where generation and consumption are jointly optimized.

To provide stronger support for the claims of faster convergence and improved stability, a quantitative comparison is added based on multiple independent training runs. Rather than relying solely on graphical learning curves, TABLE III reports convergence and stability indicators, including the number of episodes required to reach a predefined performance threshold and the variability of the final performance over the last training window. As shown, the proposed MFMC-RL method reaches the target performance in fewer episodes and exhibits lower variability than the benchmark methods, indicating both faster convergence and more stable learning behavior.

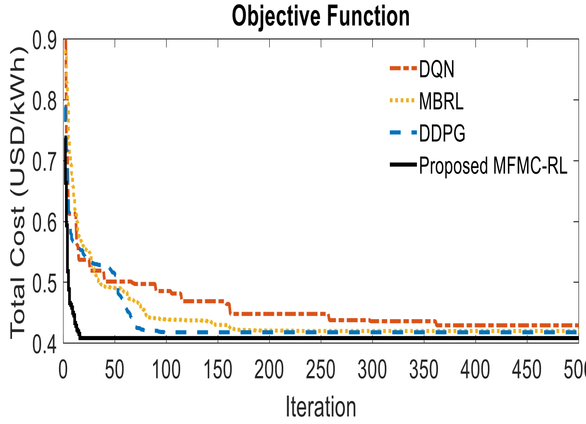

Fig. 11. Comparison of convergence rates across MFMC-RL, DDPG, DQN, and MBRL algorithms in Scenario 2.

TABLE III
Convergence Speed and Stability Metrics Across Multiple Runs

Method	Episodes to reach Convergence Threshold ($Cost \leq 0.43$)	Final Cost (Mean of last 50 episodes)	Stability Metric (Std. Dev. of last 50 episodes)
DDPG	85 ± 12	0.418 ± 0.005	0.0045
MBRL	175 ± 18	0.421 ± 0.007	0.0062
DQN	260 ± 25	0.432 ± 0.009	0.0088
Proposed MFMC-RL	**22 ± 3**	**0.410 ± 0.002**	**0.0012**

As presented in TABLE III, the quantitative analysis confirms the superiority of the MFMC-RL framework. The proposed method reaches the convergence threshold in only 22 episodes on average, approximately 3.8 times faster than DDPG and 11.8 times faster than DQN. Furthermore, the stability metric (measured as the standard deviation of the total cost during the final 50 training episodes) for MFMC-RL is 0.0012, significantly lower than the benchmark algorithms. This indicates that the Markov Chain-guided mechanism not only accelerates the search for the optimal policy but also effectively suppresses the oscillations commonly observed in standard RL training.

Fig. 12 illustrates the optimal dispatch profiles of the dispatchable generating units obtained under all scenarios. The results demonstrate the algorithm's ability to dynamically allocate generation output in response to fluctuating electrical and thermal loads. Notably, the MFMC-RL algorithm successfully coordinates the dispatch of these units to minimize total fuel consumption while strictly adhering to the combined heat and power (CHP) feasibility curves. The power outputs exhibit smooth transitions and precise tracking of load demand, indicating that the Markov Chain-guided regularization prevents the erratic fluctuations often observed in standard Q-learning, thereby ensuring stable operation and extending the lifespan of the generating assets.

The dispatch profiles for the environmentally dependent resources, the PV System, and the Wind Turbine, exhibit near-perfect alignment across all three scenarios. This consistency is fundamental, as the power output from these sources is strictly dictated by exogenous factors (solar irradiance and wind speed profiles) and is therefore not subject to the agents' control or optimization strategy. The resulting dispatch shapes for PV (a clear midday peak) and Wind (a localized afternoon peak) serve as the fixed input foundation upon which all operational decisions in Scenarios 1, 2, and 3 are based.

The primary differentiation in operational strategy is evident in the dispatchable units: the Microturbine, Fuel Cell, Diesel Generator, and the Energy Storage System. In Scenario 2, the generators actively learn the optimal policy, which leads to a more assertive and specialized utilization pattern. For instance, the FC (e) maintains a near-constant high output after 8:00 in Scenario 2, potentially taking over a baseline role that is instead shared more dynamically with the grid or ESS in Scenarios 1 and 3. This suggests the learned policies for generation units prioritize their operating envelopes to achieve the defined cost objective when given control.

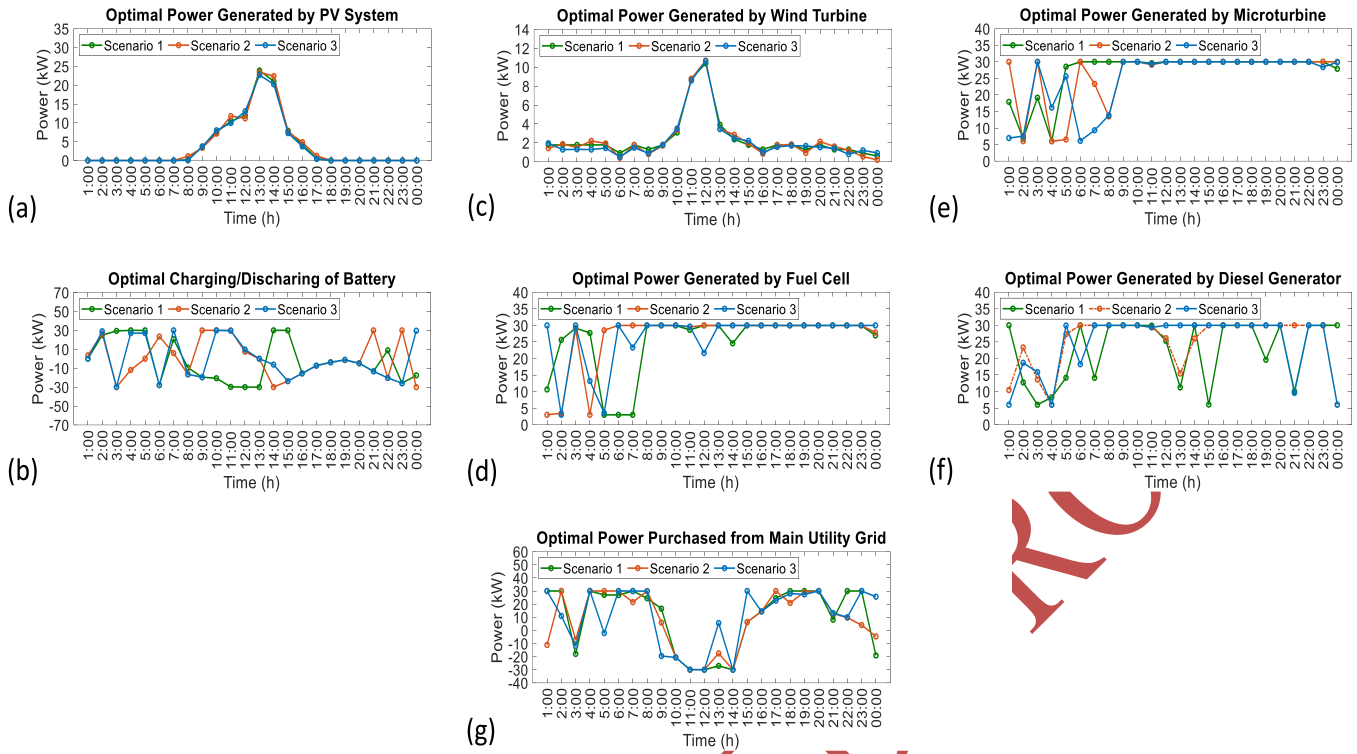

Fig. 12. Optimal Power Dispatch Profiles for the Hybrid Microgrid Components across Learning Scenarios (MFMC-RL Execution), (a) Optimal Power Generated by PV System; (b) Optimal Charging/Discharging of Battery (Positive = Charging, Negative = Discharging); (c) Optimal Power Generated by Wind Turbine; (d) Optimal Power Generated by Microturbine; (e) Optimal Power Generated by Fuel Cell; (f) Optimal Power Generated by Diesel Generator; and (g) Optimal Power Purchased from Main Utility Grid.

B. Sensitivity Analysis

To assess the robustness of the proposed MFMC-RL framework, we conducted a sensitivity analysis by varying (i) renewable generation volatility, (ii) load uncertainty, and (iii) battery energy storage constraints. For renewable and load uncertainty, Gaussian perturbations were applied to the baseline PV/wind and demand profiles with different uncertainty levels. For battery sensitivity, the storage capacity and charge/discharge limits were tightened to represent reduced flexibility.

TABLES IV and V report the average operational cost, convergence speed (episodes to reach a predefined threshold), and a stability indicator (standard deviation of the objective over the final training window), all computed over multiple independent runs (different random seeds). Overall, MFMC-RL consistently maintains lower cost and lower variability compared with DDPG, DQN, and MBRL. As uncertainty increases or storage flexibility is restricted, all methods experience some degradation; however, MFMC-RL shows the smallest degradation in both convergence speed and stability, confirming the benefit of the Markov Chain-guided update mechanism.

As demonstrated in TABLES IV and V, the sensitivity analysis confirms that the proposed MFMC-RL framework is more resilient to environmental and operational uncertainties. In the ‘High Uncertainty’ scenario (TABLE IV), the operational cost for MFMC-RL increases by only 2.9%, whereas DDPG and DQN experience increases of 5.5% and 9.2%, respectively. Furthermore, TABLE V indicates that as the energy storage capacity is restricted to 50% of its baseline, MFMC-RL maintains a significantly lower constraint violation rate (0.42%) compared to DQN (3.40%). This robustness is attributed to the Markov Chain integration, which provides a more stable probabilistic mapping of the state transitions, even when the environment becomes highly stochastic or resources are limited.

TABLE VI
Sensitivity Analysis Under Uncertainty in Renewables and Load

Uncertainty Level	Method	Avg. Operational Cost (USD/kWh)	Episodes to Convergence	Stability (Final-window Std.)
Low (Baseline)	**Proposed MFMC-RL**	**0.410 ± 0.002**	**22 ± 3**	**0.0012**
	DDPG	0.418 ± 0.005	85 ± 12	0.0045
	MBRL	0.421 ± 0.007	175 ± 18	0.0062
	DQN	0.432 ± 0.009	260 ± 25	0.0088
Medium (Uncertainty +15%)	**Proposed MFMC-RL**	**0.415 ± 0.004**	**28 ± 5**	**0.0018**
	DDPG	0.426 ± 0.008	105 ± 15	0.0072
	MBRL	0.432 ± 0.011	210 ± 22	0.0095
	DQN	0.448 ± 0.014	310 ± 35	0.0120

High (Uncertainty +30%)	**Proposed MFMC-RL**	**0.422 ± 0.007**	**35 ± 8**	**0.0031**
	DDPG	0.441 ± 0.012	140 ± 20	0.0115
	MBRL	0.455 ± 0.015	280 ± 30	0.0155
	DQN	0.472 ± 0.021	380 ± 45	0.0195

TABLE V
Sensitivity Analysis Under Battery Capacity Constraints

Battery Capacity (E_{max})	Method	Avg. Operational Cost (USD/kWh)	Constraint Violation Rate (%)
100% (Full)	**Proposed MFMC-RL**	**0.410 ± 0.002**	**0.05%**
	DDPG	0.418 ± 0.005	0.12%
	DQN	0.432 ± 0.009	0.45%
70% (Reduced)	**Proposed MFMC-RL**	**0.428 ± 0.005**	**0.18%**
	DDPG	0.442 ± 0.009	0.65%
	DQN	0.465 ± 0.015	1.12%
50% (Strict)	**Proposed MFMC-RL**	**0.445 ± 0.009**	**0.42%**
	DDPG	0.471 ± 0.014	1.85%
	DQN	0.510 ± 0.022	3.40%

V. Conclusions

This study addresses the challenging problem of energy management in hybrid microgrids by proposing a novel Model-Free Markov Chain-based Reinforcement Learning (MFMC-RL) framework. By combining the probabilistic structure of Markov Chains with the adaptability of model-free reinforcement learning, the proposed approach reduces learning instability. It mitigates the computational and tuning burdens commonly observed in centralized optimization and conventional RL-based dispatch strategies. Extensive simulations using real-world data from Iranian renewable energy installations demonstrate that MFMC-RL delivers superior and more reliable performance than state-of-the-art baselines, including DDPG, DQN, and a representative model-based RL method.

In particular, MFMC-RL achieves faster and more stable learning behavior and improves robustness under uncertainty. As quantified in TABLE IV, under the High Uncertainty setting, MFMC-RL shows only a 2.9% increase in operational cost, compared to 5.5% for DDPG and 9.2% for DQN. Furthermore, when operational flexibility is reduced by limiting the energy storage capacity to 50% of its nominal value, MFMC-RL maintains a low constraint violation rate of 0.42%, substantially below DQN (3.40%) (see TABLE V). These results indicate that the Markov Chain–enhanced update improves the stability of the learned policy by providing a more consistent probabilistic mapping of state transitions, even in highly stochastic environments.

Among the studied operating modes, the Full Learning scenario—where both generation- and demand-side agents are jointly optimized—achieves the best overall performance. This finding confirms that coordinated control of dispatchable units while explicitly considering coupled thermal and electrical constraints is essential for reducing fuel usage, lowering grid dependency, and improving operational economy in hybrid microgrids.

Despite the promising outcomes, several limitations remain. (i) The current implementation is validated on a single hybrid microgrid and does not explicitly address coordinated operation across multiple microgrids. (ii) The Markov Chain model primarily captures short-term stochastic transitions; capturing longer temporal dependencies may require higher-order Markov models or sequence-based function approximators (e.g., recurrent or transformer-based representations). (iii) The computational and memory requirements may grow with increased state dimensionality and a larger set of controllable assets and uncertainty sources, which may necessitate further scalability enhancements.

Future work will focus on extending MFMC-RL to coordinated multi-microgrid operation, leveraging transfer learning to accelerate adaptation across different geographic locations and seasonal regimes, and investigating richer state representations and scalable training schemes to further improve generalization and computational efficiency.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Authors Contribution Statement

Ehsan Akbari: Conceptualization; Methodology; Formal Analysis; Numerical Implementation; Writing—original draft, Project administration; Investigation; Validation; Writing—review & editing.

AI disclosure

The authors used ChatGPT (OpenAI, San Francisco, USA) to assist in language editing and improving the clarity of the manuscript. All AI-assisted outputs were reviewed and verified by the authors.

Data availability

No new data were created or analyzed during this study. Data sharing is not applicable to this article.

References

[1] K. A. Moungos, D. G. Kyriakou, and F. D. Kanellos, "Towards the integration of interconnected microgrids to deregulated electricity markets," *Electric Power Systems Research*, vol. 239, pp. 111264–111264, Nov. 2024.

[2] Q. Hassan, S. Algburi, A. Z. Sameen, H. M. Salman, and M. Jaszczur, "A Review of Hybrid Renewable Energy systems: Solar and wind-powered solutions: Challenges, opportunities, and Policy Implications," *Results in Engineering*, vol. 20, no. 1, p. 101621, Nov. 2023.

[3] M. A. Kamarposhti, I. Colak, H. Shokouhandeh, C. Iwendi, Sanjeevikumar Padmanaban, and S. S. Band, "Optimum operation management of microgrids with cost and environment pollution reduction approach considering uncertainty using multi-objective NSGAII algorithm," *IET renewable power generation*, vol. 19, no. 1, Aug. 2022.

[4] A. Dong and S.-K. Lee, “The Study of an Improved Particle Swarm Optimization Algorithm Applied to Economic Dispatch in Microgrids,” *Electronics*, vol. 13, no. 20, pp. 4086–4086, Oct. 2024.

[5] M. Naguib, W. A. Omran, and H. E. A. Talaat, “Performance Enhancement of Distribution Systems via Distribution Network Reconfiguration and Distributed Generator Allocation Considering Uncertain Environment,” *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 3, pp. 647–655, May 2022.

[6] M. Alilou, B. Tousi, and H. Shayeghi, “Home energy management in a residential smart micro grid under stochastic penetration of solar panels and electric vehicles,” *Solar Energy*, vol. 212, pp. 6–18, Dec. 2020.

[7] M. Amini, A. Khorsandi, B. Vahidi, S. H. Hosseinian, and A. Malakmahmoudi, “Optimal sizing of battery energy storage in a microgrid considering capacity degradation and replacement year,” *Electric Power Systems Research*, vol. 195, p. 107170, Jun. 2021.

[8] A. Sui, F. Wei, X. Lin, C. Wu, Z. Wang, Z. Li, “Multi-energy-storage energy management with the robust method for distribution networks,” *Electrical Power and Energy System*, vol. 118, pp. 1-15, June 2020.

[9] M. Cao, Q. Xu, X. Qin, J. Cai, “Battery energy storage sizing based on a model predictive control strategy with operational constraints to smooth the wind power,” *Electrical Power and Energy Systems*, vol. 115, pp. 1-10, Feb. 2020.

[10] L. Lei, Y. Tan, G. Dahlenburg, W. Xiang, and K. Zheng, "Dynamic Energy Dispatch Based on Deep Reinforcement Learning in IoT-Driven Smart Isolated Microgrids," *IEEE Internet of Things Journal*, 2020.

[11] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A Multi-agent Reinforcement Learning based Data-driven Method for Home Energy Management," *IEEE Transactions on Smart Grid*, 2020.

[12] Z. Hu, M. Zhu, P. Chen, and P. Liu, "On convergence rates of game theoretic reinforcement learning algorithms," *Automatica*, Vol. 104, pp. 90-101, 2019.

[13] K. Deshpande, P. Möhl, A. Hämmerle, G. Weichhart, H. Zörrer, and A. Pichler, "Energy Management Simulation with Multi-Agent Reinforcement Learning: An Approach to Achieve Reliability and Resilience," *Energies*, Vol. 15, No. 19, p. 7381, 2022.

[14] V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529-533, 2015.

[15] T. P. Lillicrap et al., “Continuous control with deep reinforcement learning,” in *International Conference on Learning Representations* (ICLR), 2016.

[16] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 6379-6390, 2017.

[17] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Chen, and F. Blaabjerg, “Multi-agent deep reinforcement learning for microgrid energy management in smart cities,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 11, pp. 7020-7031, 2020.

[18] J. R. Vazquez-Canteli and Z. Nagy, “Reinforcement learning for demand response: A review of algorithms and modeling techniques,” *Applied Energy*, vol. 235, pp. 1072-1089, 2019.

[19] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, “Deep reinforcement learning for pricing-based demand response of smart buildings,” *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 710-721, 2020.

[20] H. Li, Z. Wan, and H. He, “Real-time energy management of a multi-energy microgrid with integrated deep reinforcement learning,” *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4190-4201, 2020.

[21] J. Lei, Y. Yang, D. Zheng, Y. Li, and C. Chen, “Deep reinforcement learning-based energy management of a multi-energy microgrid,” *Energy*, vol. 228, p. 120610, 2021.

[22] M. Daneshvar, B. Mohammadi-Ivatloo, K. Zare, and S. Asadi, “Optimal stochastic energy management of multi-microgrids with high penetration of renewable energy sources,” *International Journal of Electrical Power & Energy Systems*, vol. 121, p. 105930, 2020.

BIOGRAPHIES

Ehsan Akbari received the B.Sc. degree in Electrical Power Engineering from Mazandaran University, Babolsar, Iran, in 2009 and M.S. degree in Electrical Power Engineering from Mazandaran University of Science and Technology, Babol, Iran, in 2014. He received Ph.D. in Electrical Power Engineering from Isfahan University of Technology, Isfahan, Iran in 2022. He is now a Assistant Professor at Department of Electrical Engineering, Mazandaran University of Science and Technology, Babol, Iran. He is the author of 25 books and more than 355 papers in reputed journals and conferences and won six patents in his research fields. He has obtained five provincial scientific and technological progress awards. His main areas of research are power quality, flexible AC transmission systems (FACTS), application of power electronics in power systems, power electronics multilevel converters, smart grids, control of grid-connected converters, fault location, distributed generation, energy storage systems, micro-grids, voltage stability, electrical machines, special electrical machinery, HVDC systems, harmonics, reactive power control using hybrid filters and renewable energy systems.