

Article

A Soft Actor-Critic-Based Energy Management Strategy for Fuel Cell Vehicles Considering Fuel Cell Degradation

Handong Zeng ¹, Changqing Du ^{1,2,*}  and Yifeng Hu ¹ 

¹ Hubei Key Laboratory of Advanced Technology for Automotive Components, Wuhan University of Technology, Wuhan 430070, China; 295286@whut.edu.cn (H.Z.); hu_yifeng@whut.edu.cn (Y.H.)

² National Energy Key Laboratory for New Hydrogen-Ammonia Energy Technologies, Foshan Xianhu Laboratory, Foshan 528200, China

* Correspondence: cq_du@whut.edu.cn

Abstract

Energy management strategies (EMSs) play a critical role in improving both the efficiency and durability of fuel cell electric vehicles (FCEVs). To overcome the limited adaptability and insufficient durability consideration of existing deep reinforcement learning-based EMSs, this study develops a degradation-aware energy management strategy based on the Soft Actor–Critic (SAC) algorithm. By leveraging SAC’s maximum-entropy framework, the proposed method enhances exploration efficiency and avoids premature convergence to operating patterns that are unfavorable to fuel cell durability. A reward function explicitly penalizing hydrogen consumption, power fluctuation, and degradation-related operating behaviors is designed, and the influences of reward weighting and key hyperparameters on learning stability and performance are systematically analyzed. The proposed SAC-based EMS is evaluated against Deep Q-Network (DQN) and Proximal Policy Optimization (PPO) strategies under both training and unseen driving cycles. Simulation results demonstrate that SAC achieves a superior and robust trade-off between hydrogen economy and degradation mitigation, maintaining improved adaptability and durability under varying operating conditions. These findings indicate that integrating degradation awareness with entropy-regularized reinforcement learning provides an effective framework for practical EMS design in FCEVs.

Keywords: fuel cell hybrid electric vehicle; fuel cell health; energy management; SAC reinforcement learning

1. Introduction

With the intensifying global energy crisis and environmental challenges, issues such as global warming, air pollution, and energy depletion have become increasingly severe, compelling society to pursue sustainable and clean energy alternatives [1]. In this context, Fuel Cell Electric Vehicles (FCEVs) have emerged as a promising pathway toward low-carbon transportation due to their zero-emission characteristics, high energy conversion efficiency, and renewable potential [2]. However, their practical implementation still faces several engineering challenges, including the relatively slow dynamic response of fuel cells [3], degradation under frequent load fluctuations, and limitations in on-board hydrogen storage efficiency [4]. Consequently, integrating fuel cells with auxiliary energy sources such as lithium-ion batteries or supercapacitors to form hybrid architectures has become a mainstream strategy to enhance power performance and system reliability in FCEVs [5].



Academic Editor: Branislav Hredzak

Received: 28 November 2025

Revised: 7 January 2026

Accepted: 13 January 2026

Published: 15 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

The architecture of FCEVs merely lays the foundation for addressing the aforementioned issues. The essential challenge lies in how to dynamically coordinate the power distribution among multiple energy sources—meeting the vehicle's power demands while simultaneously optimizing hydrogen consumption and extending the lifespan of key components. This has become a critical barrier to the large-scale commercialization of FCEVs [6]. To overcome this challenge, numerous institutions and research teams have conducted extensive studies on Energy Management Strategies (EMSs), which can be broadly classified into three categories: rule-based, optimization-based, and learning-based approaches [7].

Conventional energy management strategies are generally classified into rule-based and optimization-based methods [8]. Rule-based approaches, including fuzzy logic control [9] and deterministic logic schemes such as power-following strategies [10], rely on predefined “if-then” rules derived from prior knowledge and engineering experience. They provide strong real-time performance due to their simplicity but depend heavily on expert-defined logic, leading to limited adaptability under complex driving conditions and potential fuel cell load fluctuations that accelerate system degradation [11,12]. Optimization-based methods encompass global optimization techniques such as Dynamic Programming (DP) [13] and Pontryagin's Minimum Principle (PMP) [14], as well as instantaneous optimization strategies like EMCS [15] and Model Predictive Control (MPC) [16]. These methods formulate energy management as a constrained optimization problem by constructing accurate system models and defining appropriate cost functions. While global optimization can yield theoretically optimal solutions, its reliance on complete prior driving-cycle information and high computational burden restricts its feasibility for real-time application [17].

Reinforcement learning (RL) provides a model-free framework well suited to the complex, nonlinear, uncertain, and multi-objective characteristics of fuel cell hybrid systems and is commonly classified into value-based, policy-based, and actor-critic methods [18]. Value-based approaches, such as Q-learning and Deep Q-Networks (DQN), effectively address discrete decision-making and high-dimensional state spaces. Wang et al. [19] proposed an improved DQN-based energy management strategy that incorporates a data-driven battery lifetime map to characterize nonlinear aging and employs a parameterized DQN to handle hybrid discrete-continuous actions, achieving 99.5% of dynamic programming optimality and reducing operating costs by 3.1% under unknown conditions, while Wang et al. [20] further enhanced stability and convergence speed through a Dueling DQN architecture. Policy-based methods, including Proximal Policy Optimization (PPO) [21], directly optimize policy parameters using clipped surrogate objectives to improve training stability; Li et al. [22] integrated PPO with dynamic programming prior knowledge and parallel computation, resulting in improved convergence speed, hydrogen consumption, and fuel cell degradation compared with DQN- and DDPG-based EMS, and Lv et al. [23] developed an improved PPO-based hierarchical energy management strategy with adaptive driving modes and offline-optimized battery SOC references, achieving a 3.11% improvement in economic performance. Actor-critic methods combine value estimation and policy learning to balance exploration and exploitation; Deep Deterministic Policy Gradient (DDPG) improves stability in continuous control via target networks and delayed updates [24], while Soft Actor-Critic (SAC) further enhances exploration efficiency through maximum entropy learning [25]. Lu et al. [26] proposed an improved DDPG-based strategy incorporating adaptive fuzzy filtering for frequency-decoupled power allocation and battery degradation modeling, achieving a 2.02% increase in average fuel cell efficiency and a 14.4% reduction in battery performance degradation compared with conventional DDPG.

Despite these advances, significant gaps remain in comparative analyses of RL-based EMS. Few studies systematically compare RL algorithms across different architectures and

investigate the impact of hyperparameters and reward weights on system performance. Wang et al. [27] designed and compared four types of EMS while considering parameter effects on optimization performance, but their work focused on gasoline-electric hybrid systems. Xu et al. [28] designed and compared three types of RL-based EMS algorithms but did not discuss the effects of algorithm parameters and reward functions on optimization performance. Therefore, comprehensive research on the impacts of algorithm parameters, reward design, and hyperparameter tuning is crucial. Such studies not only deepen our understanding of RL's potential in the EMS domain but also provide practical guidance for optimizing reward shaping, algorithm selection, and parameter tuning. This work holds both theoretical and practical significance, particularly in optimizing RL-based EMS under complex real-world driving conditions. It will advance the efficiency, reliability, and service life of fuel cell hybrid vehicles.

To address the aforementioned challenges, this study first establishes a dynamic simulation model of a fuel cell hybrid electric vehicle (FCHEV). Building on this model, a degradation-aware energy management strategy based on the Soft Actor–Critic (SAC) algorithm is developed. By exploiting SAC's entropy-regularized learning mechanism, the proposed strategy explicitly balances hydrogen economy, power smoothness, and fuel cell degradation during long-term operation. The influences of key algorithmic hyperparameters and reward-weight configurations on training stability and control performance are systematically investigated. Furthermore, the proposed method is benchmarked against power-following, DQN-based, and PPO-based strategies in terms of tuning complexity, training efficiency, performance under training conditions, and generalization capability under unseen driving cycles. Through these analyses, this work provides a comprehensive assessment of the suitability of different reinforcement learning paradigms for energy management in fuel cell hybrid vehicles. In summary, the main contributions of this work are as follows:

1. A degradation-aware SAC-based energy management strategy is proposed, in which fuel cell durability and hydrogen economy are jointly incorporated into the reward function. Leveraging SAC's maximum-entropy formulation, the strategy promotes operation within high-efficiency regions while mitigating degradation-inducing power fluctuations.
2. A systematic sensitivity analysis of SAC hyperparameters and reward design is conducted, revealing their respective roles in convergence behavior, training stability, and energy management performance. The results provide practical guidelines for hyperparameter tuning and reward shaping in RL-based EMS design.
3. A comprehensive comparative evaluation of power-following and RL-based EMSs (SAC, DQN, and PPO) is performed with respect to durability, economic performance, adaptability to unseen driving cycles, tuning complexity, and training efficiency. The comparative results clarify the strengths, limitations, and applicable scenarios of different RL algorithm categories for fuel cell hybrid vehicle energy management.

The remainder of this paper is organized as follows: Section 2 presents the powertrain architecture of the fuel cell hybrid vehicle and the modeling of its key components. Section 3 introduces the proposed SAC-based energy management strategy that considers fuel cell degradation and investigates the effects of hyperparameters and reward function weightings on convergence behavior and learning performance. Section 4 compares the training outcomes and optimization performance of the three different algorithms. Finally, Section 5 summarizes and discusses the entire study and presents concluding remarks.

2. Vehicle Model and Architecture

Based on the electric–electric-hybrid powertrain architecture, a multi-energy system model is established, consisting of a Fuel Cell (FC), a Lithium Battery (LB), a DC/DC converter, and a traction load, as illustrated in Figure 1. In this configuration, the fuel cell is connected to the DC bus via a unidirectional DC/DC converter, while the lithium battery is directly coupled to the bus, jointly supplying the power demanded by the drive motor [29]. In the proposed system, the lithium battery is passively connected to the DC bus without a bidirectional DC–DC converter and therefore is not actively controlled. Its charging and discharging behavior is determined by the DC bus voltage, the battery open-circuit voltage, and internal resistance, allowing it to act as a passive energy buffer for transient power compensation. The DC bus voltage is regulated by the fuel cell DC–DC converter operating in voltage control mode, while the lithium battery indirectly contributes to voltage stabilization by smoothing power fluctuations. The modeling process primarily involves the following key components:

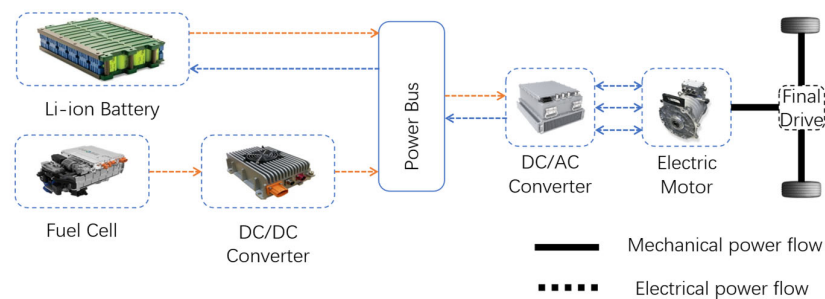


Figure 1. Power System Architecture Diagram.

2.1. Vehicle Dynamics Model

In the vehicle dynamics model, all power sources are integrated and transformed into the vehicle's motion. As shown in Equation (1), the required driving force F_{req} of the vehicle is provided by the electric motor, while the corresponding resistive forces include the acceleration resistance F_a , rolling resistance F_r , road gradient resistance F_i , and aerodynamic drag F_w . The total driving resistance F_f represents the combined effect of rolling resistance F_r and aerodynamic drag F_w [30]. The relationship can be expressed as follows:

$$\begin{cases} P_{req} = v \cdot F_{req} \\ F_{req} = F_a + F_r + F_i + F_w \\ F_a = m \cdot a \\ F_r = \mu mg \cos \theta \\ F_i = mg \sin \theta \\ F_w = A_w C_d \frac{v^2}{21.15} \\ F_f = F_a + F_r \\ F_f = A + Bv + Cv^2 \end{cases} \quad (1)$$

where μ denotes the rolling resistance coefficient, v represents the vehicle speed, A_w is the frontal area of the vehicle, C_d is the aerodynamic drag coefficient, θ denotes the road gradient, and g is the gravitational acceleration, taken as 9.8 m/s^2 .

The actual motor power, rotational speed, torque, and the resulting electrical power can be expressed as follows:

$$T_m = \frac{T_w}{i_w \eta_w} \quad (2)$$

$$n_m = n_w i_w \quad (3)$$

$$P_e = \frac{9550T_m}{n_m\eta_m} \quad (4)$$

where T_m denotes the motor torque, and T_w represents the wheel torque. n_m and n_w indicate the rotational speeds of the motor and the wheels, respectively. i_w is the main transmission ratio, η_w represents the transmission efficiency, and η_m denotes the motor efficiency. The vehicle parameters and motor efficiency map are presented in Table 1 and Figure 2, respectively.

Table 1. Vehicle parameters.

System	Item	Value & Unit
Vehicle	Vehicle mass m_v	2334 kg
	Drag coefficient A A_f	103.9 N
	Drag coefficient B C_d	2.1515 Nh/km
	Drag coefficient C ρ	0.0234 Nh ² /km ²
	Main transmission ratio i_0	12.29
	Transmission efficiency η_t	0.92
	Tire rolling radius r_t	342 mm
Lithium battery	Shaft moment of inertia J_s	2.5 kgm ²
	Battery capacity Q_b	50 Ah
Motor	Operating voltage U_b	390 V
	Maximum power P_{max}	200 kW
	Maximum speed ω_{max}	16,000 rpm
	Maximum torque T_{max}	311 Nm

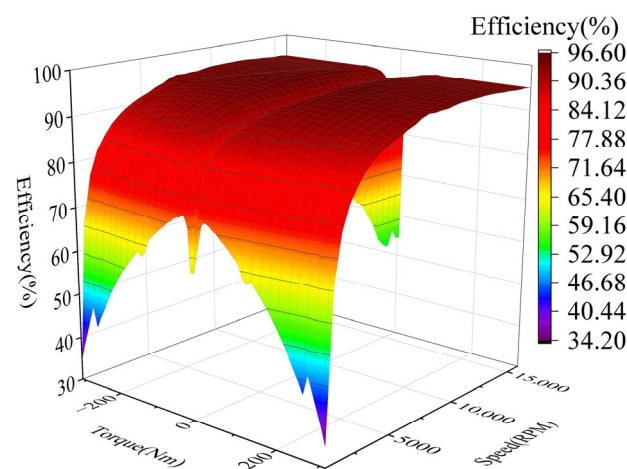


Figure 2. Motor efficiency map.

2.2. Fuel Cell System Model

A fuel cell stack consists of multiple single cells, whose output characteristics are described by an electrochemical polarization model. The output voltage U_{fc} of each cell is determined by the Nernst potential together with the activation, ohmic, and concentration overpotentials [31].

$$U_{fc} = E_{ernst} - V_{act} - V_{ohm} - V_{conc} \quad (5)$$

where E_{ernst} denotes the reversible potential, whereas V_{act} , V_{ohm} , and V_{conc} represent the activation, ohmic, and concentration losses, respectively.

$$E_{ernst} = \frac{\Delta G}{2F} - \frac{\Delta S(T - T_{ref})}{2F} + \frac{RT(\ln P_{H_2} + 0.5P_{O_2})}{2F} \quad (6)$$

where ΔG is the Gibbs free energy, F is the Faraday constant, ΔS is the entropy change in the system, T is the system temperature, T_{ref} is the reference temperature, R is the universal gas constant, P_{H_2} is the partial pressure of hydrogen, and P_{O_2} is the partial pressure of oxygen.

$$V_{act} = \xi_1 + \xi_2 T + \xi_3 T \ln C_{O_2} + \xi_4 \ln I_{st} \quad (7)$$

where ξ_1 , ξ_2 , ξ_3 , and ξ_4 are coefficients affected by the system temperature and pressure, C_{O_2} is the molar concentration of oxygen, and I_{st} is the fuel cell current.

$$V_{ohm} = I_{st} R_{int} = I_{st} (R_M + R_C) \quad (8)$$

where R_{int} is the equivalent internal resistance, R_M is the equivalent membrane resistance associated with proton conduction, and R_C is the equivalent contact resistance associated with electron conduction.

$$V_{conc} = -B \ln \left(1 - \frac{J}{J_{max}} \right) \quad (9)$$

The dynamic response behavior of the fuel cell is approximated by a first-order inertial element, wherein the output power (P_{fc}) conforms to the following relationship:

$$\tau \frac{dP_{fc}}{dt} + P_{fc} = P_{fc,cmd} \quad (10)$$

where τ denotes the time constant, taken as 2 s, and $P_{fc,cmd}$ represents the commanded power, which is constrained within the maximum and minimum operating limits of the fuel cell.

From the perspective of overall vehicle economy, it is also essential to account for the fuel cell efficiency and hydrogen consumption. The hydrogen consumption rate can be expressed as follows:

$$\dot{m}_{H_2} = \frac{P_{fc}}{\eta_{fc} LHV_{H_2}} \quad (11)$$

where P_{fc} denotes the fuel cell power, η_{fc} represents the fuel cell operating efficiency, and LHV_{H_2} refers to the lower heating value of hydrogen, which is taken as 120 MJ/kg. Accordingly, the fuel cell efficiency can be expressed as follows:

$$\eta_{fc} = \frac{P_{fc}}{\dot{m}_{H_2} \cdot LHV_{H_2}} \quad (12)$$

The specific relationships among fuel cell power, hydrogen consumption rate, and operating efficiency are illustrated in Figure 3.

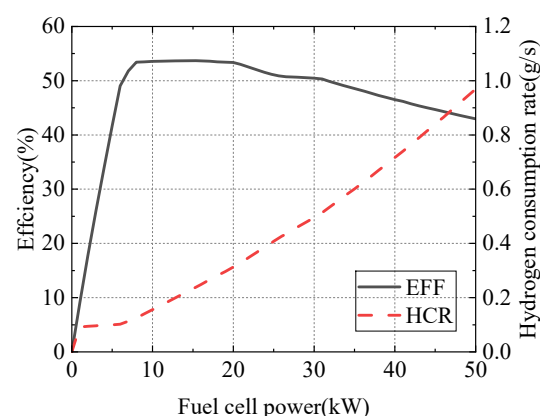


Figure 3. Fuel cell efficiency and hydrogen consumption rate.

2.3. Fuel Cell Degradation Model

During the operation of a fuel cell, four primary operating conditions contribute to voltage degradation and consequently affect its lifespan: low-power operation, high-power operation, load fluctuation, and start–stop cycling [32]. Low-power operation may lead to increased cathode potential and local fuel starvation, accelerating carbon support corrosion and catalyst degradation. High-power operation results in high current density and elevated temperature, which intensify membrane dehydration, catalyst sintering, and electrochemical aging. Load fluctuations induce frequent transient changes in voltage and current, causing repeated mechanical and chemical stresses on the membrane–electrode assembly, thereby accelerating material fatigue. In addition, start–stop cycling leads to large potential excursions and reactant redistribution, which promote carbon corrosion and irreversible performance loss. Therefore, these operating conditions are widely recognized as the dominant contributors to fuel cell lifespan degradation and are incorporated into the degradation model in this study. In this study, the fuel cell is assumed to remain continuously active; thus, the impact of start–stop conditions on degradation is neglected. Accordingly, the empirical expression for the instantaneous degradation rate ($\dot{D}$) of the fuel cell is given as follows [33]:

$$\dot{D} = k_1 t_1 + k_2 t_2 + k_3 \Delta P_{fc} \quad (13)$$

where k_1 , k_2 , and k_3 represent the instantaneous degradation rates for the fuel cell under low-power, high-power, and load-varying conditions, respectively. t_1 and t_2 denote the total operating time under low-power and high-power conditions, respectively, and ΔP_{fc} denotes the total power change throughout the operating process. The values of these parameters are given in Table 2 [34].

Table 2. Fuel cell degradation values.

Operating Conditions	Item	Value	Unit
High power	k_1	1.26×10^{-3}	%h
Low power	k_2	1.47×10^{-3}	%h
Load varying	k_3	5.39×10^{-5}	%kWh

It should be noted that the degradation model adopted in this study is an empirical, system-level formulation intended for comparative evaluation of energy management strategies rather than precise lifetime prediction. Micro-scale mechanisms such as liquid water accumulation in the gas diffusion layer (GDL), which can restrict reactant transport under low-power operation and accelerate catalyst degradation, are not explicitly modeled, although recent studies have demonstrated that such effects can be quantified using lattice Boltzmann methods in gradient-porosity GDLs [35]. In addition, the model parameters are not calibrated using dynamic aging bench tests, nor are temperature–current density coupling effects explicitly considered, even though elevated temperatures are known to exacerbate membrane electrode degradation. Advanced online health state estimation and degradation diagnostics under dynamic loads, as well as temperature–aging coupling models reported in recent literature, provide valuable theoretical and methodological references for improving degradation modeling fidelity [36]. Integrating these mechanisms with degradation-aware reinforcement learning-based energy management remains an important direction for future work.

2.4. Lithium Battery Model

The lithium battery pack, serving as the secondary power source in the hybrid electric vehicle, is primarily used to supply peak power and store excess energy. A Rint equivalent resistance model [37] is employed to model the lithium battery pack:

$$\begin{cases} P_{bat} = V_{OC}I - R_{bat}I^2 \\ I = \frac{V_{OC} - \sqrt{V_{OC}^2 - 4R_{bat}P_{bat}}}{2R_{bat}} \\ SOC = SOC_0 - \frac{\int Idt}{C_n} \end{cases} \quad (14)$$

where V_{OC} is the open-circuit voltage, I is the output current, R_{bat} is the internal resistance, and SOC_0 is the initial state-of-charge. The characteristics of an individual cell within the lithium battery pack are described in Figure 4. The total number of cells in the entire lithium battery pack is 122.

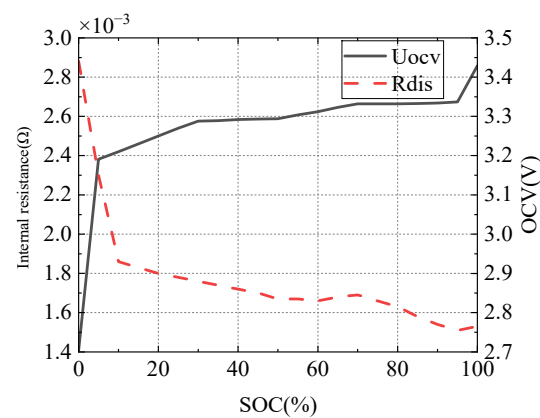


Figure 4. Characteristics of the individual cell.

2.5. Model Simulation and Verification

The aforementioned model was implemented and simulated in MATLAB/Simulink 2023a with a simulation step size of 0.1 s. All module data were derived from real vehicle testing. To validate the rationality and accuracy of the model, a comparison between the simulation results and experimental data is presented in Figure 5. From left to right: vehicle speed, fuel cell system power, and hydrogen consumption.

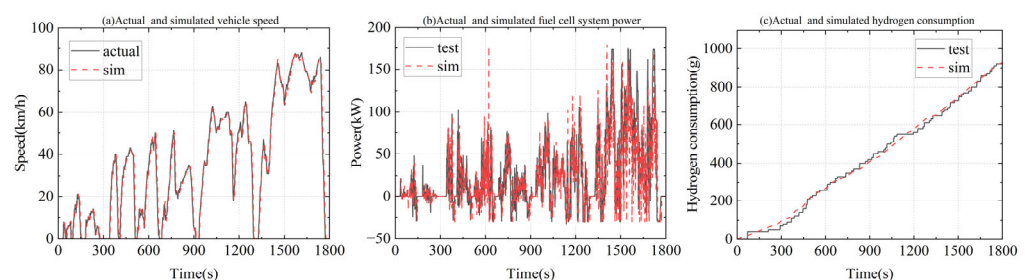


Figure 5. Simulation results and experimental data.

As can be observed from the figures, the simulation curves are closely aligned with the experimental data from the real vehicle. The Root Mean Square Error (RMSE) between the simulated and experimental values is less than 3% of the RMSE of the corresponding experimental data, which validates the model's rationality and demonstrates its close approximation to real-world conditions.

3. Degradation-Aware SAC for Energy Management

3.1. SAC Algorithm Framework

SAC (Soft Actor-Critic) is an off-policy deep reinforcement learning method. It is built upon the Actor-Critic framework, where the Actor network outputs a stochastic policy to enhance exploration. Distinguished from the original Actor-Critic architecture, the SAC agent aims to maximize the expected cumulative reward while also maximizing the entropy of the policy. The state-action value function is governed by the soft Bellman equation:

$$Q(s_t, a_t) = r_t + \gamma E_{s_{t+1}, a_{t+1}} \left[\begin{array}{c} Q(s_{t+1}, a_{t+1}) \\ -\alpha \log(\pi(a_t | s_t)) \end{array} \right] \quad (15)$$

where s_t , a_t , and r_t denote the state, action, and reward at time step t , respectively; s_{t+1} and a_{t+1} represent the state and action after state transition; γ is the discount factor; E denotes the mathematical expectation; and α is the temperature parameter that regulates the relative importance of the entropy term compared to the reward, which is automatically adjusted by the neural network.

$\pi(a_t | s_t)$ is the policy to be learned, and the optimal policy is expressed as:

$$\pi^* = \operatorname{argmax}_{\pi} \sum_t E_{(s_t, a_t)} \rho \pi [r_t - \alpha \log(\pi(a_t | s_t))] \quad (16)$$

Neural networks are employed to approximate the Q -function, and the policy can be modeled as a Gaussian distribution, whereby its mean and covariance are output by neural networks. Consequently, during the backpropagation process, both the actor and critic can be optimized via stochastic gradient descent. The parameters θ of the Q -function can be trained by minimizing the soft Bellman residual:

$$J_Q(\theta) = E_{(s_t, a_t, r_t, s_{t+1})} \frac{1}{2} \left[Q(s_t, a_t) - \left[r_t + \gamma \left(\begin{array}{c} Q'(s_{t+1}, \pi(s_{t+1})) \\ -\alpha \log(\pi(a_{t+1} | s_{t+1})) \end{array} \right) \right] \right]^2 \quad (17)$$

where M represents the experience replay buffer, and (s_t, a_t, r_t, s_{t+1}) denotes a mini batch randomly sampled from it. A target critic network Q' is utilized to accelerate and stabilize the training process, whose parameters θ' are updated via a soft update mechanism:

$$\theta' \leftarrow (1 - \tau) \theta' + \tau \theta \quad (18)$$

where τ is the step size factor that governs the magnitude of the update. The policy network parameters are updated by minimizing the following objective:

$$J_{\pi}(\phi) = E_{s_t} \left[E_{a_t} \pi_{\phi} \left[\begin{array}{c} \alpha \log(\pi(a_t | s_t)) \\ -Q(s_t, a_t) \end{array} \right] \right] \quad (19)$$

The temperature parameter α is automatically adjusted, with its gradient calculated to minimize the following objective:

$$J(\alpha) = E_{E_{a_t} \pi_t} \left[-\alpha \log \pi_t(a_t | s_t) - \alpha \bar{H} \right] \quad (20)$$

where $\bar{H}$ is the target entropy, typically set to the negative of the action dimension. The parameters of α are updated accordingly during the training process. The specific structure of the SAC algorithm is shown in Figure 6.

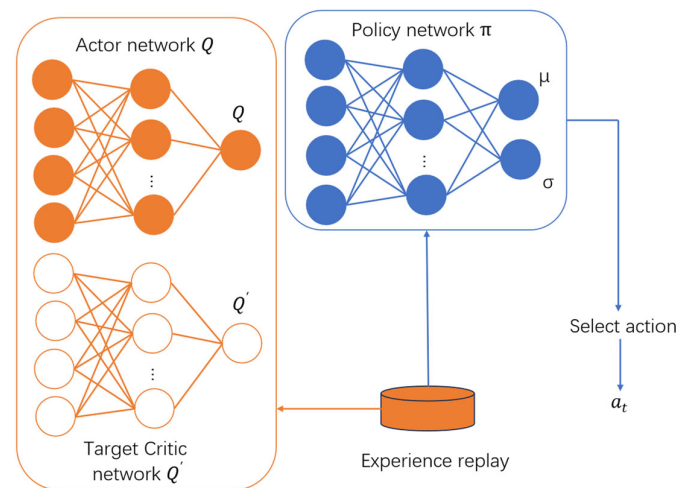


Figure 6. SAC structure.

3.2. Design Based on SAC Reinforcement Learning

3.2.1. State Space

The subject of this study is a fuel cell hybrid electric vehicle, which utilizes two power sources: a fuel cell and a lithium battery. The state space of the overall operating environment is defined as follows:

$$S = \{SOC, V, P_{dem}\}$$

where SOC denotes the lithium battery state of charge, V represents the vehicle speed, and P_{dem} indicates the total power demand of the vehicle.

3.2.2. Action Space

The action space in this study is defined as the power output of the fuel cell, while the lithium battery passively balances the total vehicle power demand:

$$P_{batt} = P_{dem} - P_{fc} \quad (21)$$

where P_{batt} and P_{fc} represent the power output of the lithium battery and the fuel cell, respectively. The fuel cell power output is designated as the action space for the reinforcement learning agent:

$$A = P_{fc}$$

where $P_{fc} \in [0, 40]$ kW. However, since the DQN reinforcement learning algorithm only supports discrete action spaces, the action space for DQN is defined as $P_{fc} = \{0, 1, 2, 3, \dots, 40\}$ kW.

3.2.3. Reward Function

The reward function primarily considers the fuel cell's power consumption, fuel consumption, and the power change rate leading to lifetime degradation. The specific reward function R is defined as follows:

$$R = -\left[\alpha \cdot \dot{m}_{H_2} + \beta \cdot |SOC(t) - SOC_{ref}| \cdot P_{H_2} + \gamma \cdot |P_{fc}(t+1) - P_{fc}(t)|\right] \quad (22)$$

where $\dot{m}_{H_2}$ is the instantaneous hydrogen consumption rate of the fuel cell, SOC_{ref} is the initial SOC value set to 0.6, and P_{H_2} represents the hydrogen consumption per unit SOC change. The weighting coefficient $\alpha = 1$ for the first term serves to determine the order of magnitude of the entire function, facilitating the narrowing of the value ranges for the subsequent weighting coefficients. The weighting coefficient $\beta = 0.049$ for the

second term functions to stabilize the lithium battery SOC and prevent rapid degradation. The weighting coefficient $\gamma = 0.1$ for the third term aims to balance the overall power fluctuations, enabling the vehicle's power sources to output power as smoothly as possible.

The weights in the reward function are selected based on a trial-and-error tuning process, with the objective of ensuring that each reward term has a comparable numerical magnitude during training. This prevents any single term from dominating the learning process and improves training stability and convergence behavior. Limited variations around the selected values were tested, and no significant performance degradation was observed.

3.2.4. Training Settings

The SAC reinforcement learning algorithm involves a relatively large number of adjustable parameters, many of which have minimal impact on the learning process and final outcomes. This paper selects the key parameters listed in the table below for design and subsequently analyzes the influence of their numerical variations on the reinforcement learning process and results. Since the SAC algorithm utilizes both actor and critic networks, the terms “number of neurons” and “learning rate” refer to the values for these two networks, respectively, as specified in Table 3.

Table 3. SAC Hyperparameters.

Item	Value
Number of neurons (Actor/Critic)	256/352
Learning rate (Actor/Critic)	$1 \times 10^{-5}/5 \times 10^{-5}$
Target entropy	−2
Smoothing factor	1×10^{-3}
Experience replay buffer size	1×10^6
Minimum batch size	64
Discount factor	0.99

It should be noted that the model employs a feedforward neural network with dual inputs, namely state and action. The actor network consists of a shared trunk and two output heads. The trunk contains five fully connected layers with a total of 192 neurons, while the output heads include three fully connected layers comprising 32 neurons. Therefore, the entire actor network contains eleven fully connected layers with a total of 256 neurons. The critic network is composed of two inputs and a shared trunk. Each input branch contains 128 neurons, and the trunk comprises five fully connected layers with a total of 96 neurons. Consequently, the critic network includes five fully connected layers with an overall total of 352 neurons.

The study adopts the CLTC-P cycle for training and the WLTC cycle for testing to evaluate the agent's transferability and robustness. Both are representative new-energy-vehicle driving cycles that cover urban, suburban, highway, and low-speed conditions and share similar overall profiles, each lasting 1800 s. CLTC-P reflects more idealized laboratory conditions, whereas WLTC is more indicative of real-world driving. This pairing allows the agent to learn from controlled data while being validated under more realistic operating conditions. The corresponding velocity and power demand profiles are shown in Figure 7.

3.3. Hyperparameter Design and Discussion

It should be noted that the results presented in this study are obtained from simulations using a single converged SAC policy under fixed system parameters and predefined driving cycles. Once the training converges, the learned policy is deterministic, and repeated simulations under identical conditions yield consistent performance outcomes.

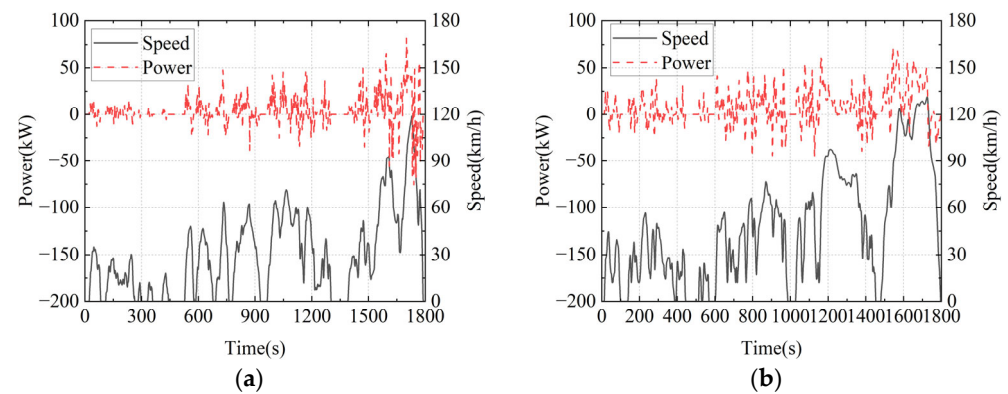


Figure 7. (a) CLTC operating condition; (b) WLTC operating condition.

3.3.1. Number of Neurons

As shown in Figure 8, the results indicate that network size does not fundamentally alter the learning dynamics of the SAC-based EMS, as all neuron configurations converge in a similar manner. However, the final performance exhibits a non-monotonic relationship with network complexity, reflecting a mismatch between representational capacity and task dimensionality. For this low-dimensional fuel cell energy management problem, a compact network is already sufficient to approximate the control policy. Introducing moderate additional complexity increases redundancy and raises the risk of convergence to suboptimal local minima. In contrast, a significantly larger network provides enhanced representational flexibility, which facilitates more effective exploration and improves the ability to escape local optima, leading to superior post-convergence performance. This finding suggests that network architecture should be selected based on task complexity rather than assuming that incremental increases in size will yield consistent performance gains.

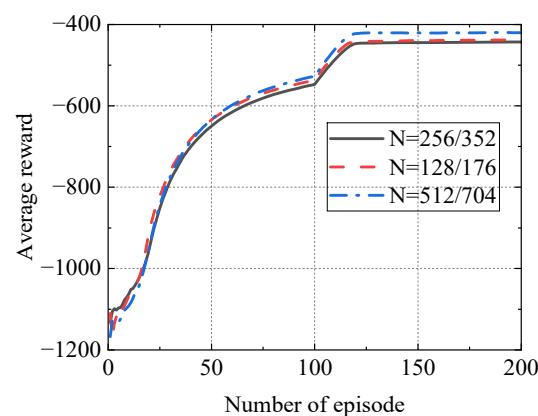


Figure 8. Effect of the number of neurons on the average reward.

3.3.2. Learning Rate

As shown in Figure 9, the learning rate significantly influences both convergence speed and asymptotic performance, reflecting the trade-off between exploration stability and optimization precision. Higher learning rates accelerate early policy improvement but increase the likelihood of premature convergence to suboptimal solutions, whereas excessively small learning rates slow down training while enabling more refined policy updates. The observed superior steady-state performance at the lowest learning rate suggests that finer gradient steps are beneficial for this control task, which features a relatively smooth and low-dimensional optimization landscape. These results indicate that aggressive learning-rate settings may sacrifice long-term optimality for faster convergence, and that learning rates below the 1×10^{-5} – 1×10^{-6} range are more suitable for balancing

stability and performance in degradation-aware fuel cell energy management. Nonetheless, identifying the globally optimal value requires more systematic hyperparameter search beyond single-factor analysis.

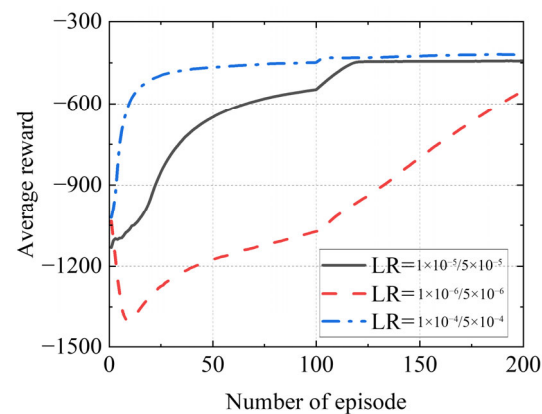


Figure 9. Effect of Learning Rate on Average Reward.

3.3.3. Target Entropy

As shown in Figure 10, the impact of target entropy on training largely aligns with theoretical expectations, with the exception of the case where target entropy is 0. In this scenario, a target entropy of 0 results in excessive exploration, which hinders convergence. For target entropy values of -1 , -2 , and -3 , all configurations successfully converge, but the extreme values of entropy—both the lowest and the highest—result in suboptimal performance compared to the intermediate value. The decline in performance at lower entropy is attributed to insufficient exploration, which leads to convergence at local optima. Conversely, the reduction in reward at higher entropy indicates either incomplete exploration or convergence to a suboptimal solution.

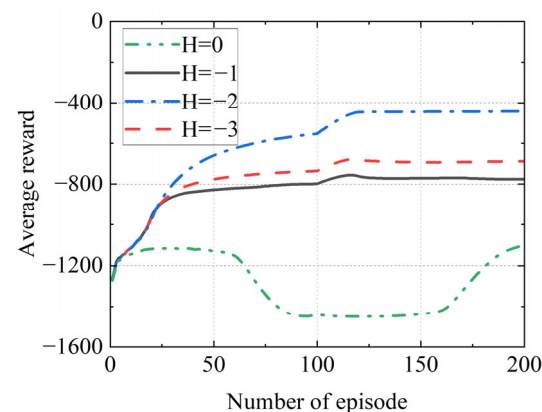


Figure 10. Effect of Target Entropy on Average Reward.

3.3.4. Reward Function Weights

To investigate the impact of each weight in the reward function on the learning outcomes, this study compares the learning processes and optimized performance of agents trained with weights halved and doubled from their initial settings against those trained with the original values. Figure 11 illustrates the influence of different weight configurations on the learning process, while Tables 4 and 5 present the driving range and fuel cell degradation rate under each weight setting (here and hereafter, the driving range refers to the distance achievable by consuming 4.2 kg of hydrogen).

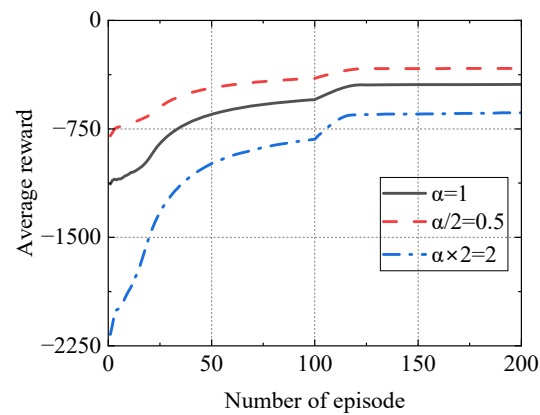


Figure 11. Average reward under different weight configurations.

Table 4. Driving range under different weight configurations (km).

Weight Item	Weight Value/2	Benchmark Weight Value	Weight Value $\times 2$
α	556.94	559.81	557.78
β	559.15	559.81	559.46
γ	556.77	559.81	555.84

Table 5. Degradation rate under different weight configurations ($\times 10^{-4}\%$).

Weight Item	Weight Value/2	Benchmark Weight Value	Weight Value $\times 2$
α	5.08	5.04	5.01
β	5.00	5.04	5.21
γ	5.06	5.04	4.97

Figure 11 highlights the effect of modifying the first weight term on the average reward function. Although changes in weight do not alter the overall convergence trend, reducing the weight leads to a decrease in the absolute value of the reward function, while increasing it results in an elevation of the absolute value. This pattern aligns with the distribution observed in the curves. Since the influence of other weight terms on the reward function follows a similar trend, results for the first weight term are presented here as representative.

Tables 4 and 5 reveal that variations in the reward-function weights have a limited impact on the driving range, with fluctuations typically confined to a 3–4 km range. The maximum driving range is consistently achieved under the baseline weighting. In contrast, the degradation rate exhibits greater sensitivity to changes in weight. Specifically, increasing the weights for the hydrogen consumption and power fluctuation terms results in reduced degradation, as these adjustments promote more efficient and stable fuel cell operation. On the other hand, increasing the weight for SOC deviation leads to higher degradation, as it forces the fuel cell to more closely track the load in order to maintain SOC, which accelerates aging.

3.4. Chapter Summary

As shown in Table 6, the number of neurons influences only the convergence level, exhibiting a general trend in which smaller networks achieve higher convergence values. The learning rate affects both convergence speed and final performance: higher values accelerate convergence and yield higher convergence levels, whereas lower values lead to slower and more gradually declining convergence behavior. For target entropy, lower values reduce the convergence level due to limited exploration, while excessively high values degrade convergence speed, stability, and final performance.

Table 6. Impact of parameters on convergence performance.

Item	Number of Neurons			Learning Rate			Target Entropy		
	Low	Mid	High	Low	Mid	High	Low	Mid	High
Convergence Speed	Mid	Mid	Mid	Slow	Mid	Fast	Mid	Mid	Slow
Convergence Stability	High	High	High	Low	High	High	High	High	Low
Convergence Position	High	Mid	Low	Low	Mid	High	Mid	High	Low

Table 7 indicates that, for driving range, all three reward-weight coefficients achieve their best performance at intermediate values, with both higher and lower settings leading to reduced range. For degradation rate, the first coefficient (hydrogen consumption penalty) and the third coefficient (power fluctuation penalty) exhibit an inverse relationship with degradation, whereas the second coefficient (SOC deviation penalty) shows a direct proportional relationship.

Table 7. Impact of reward function weights on durability economics.

Item	α			β			γ		
	Low	Mid	High	Low	Mid	High	Low	Mid	High
Driving range	Low	High	Mid	Low	High	Mid	Mid	High	Low
Degradation	High	Mid	Low	Low	Mid	High	High	Mid	Low

4. Results and Discussion

4.1. Training Performance Comparison

As shown in Figure 12 and Table 8, all three algorithms exhibit similar learning trends, with rewards initially fluctuating at low values before stabilizing. Although each algorithm ultimately converges, SAC achieves the most stable post-convergence behavior, followed by PPO, while DQN shows larger fluctuations. In terms of convergence speed, PPO converges the fastest, DQN reaches stability more slowly, and SAC converges the slowest due to its dual-network architecture and required warm-up phase. These structural differences also produce notable disparities in training efficiency: DQN’s value-based updates suffer from low data utilization, PPO’s on-policy updates enable efficient and lightweight computation, and SAC’s hybrid actor–critic formulation balances exploration and value estimation but incurs higher computational overhead and increased parameter-tuning complexity.

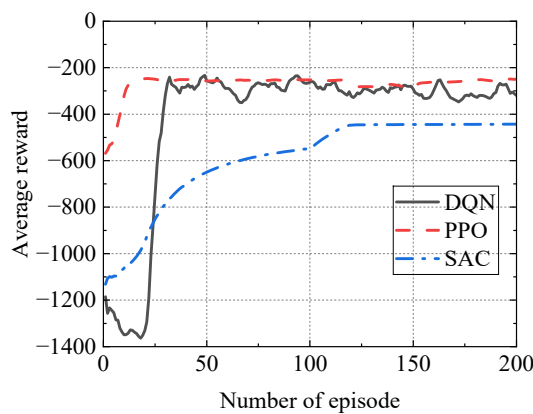


Figure 12. Average reward.

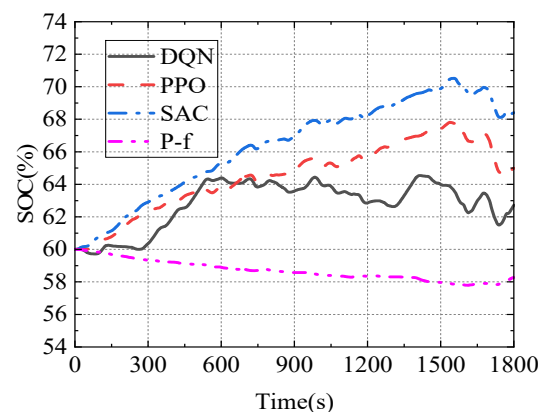
Table 8. Training Data.

Algorithm	Convergence Episodes	Training Time (s)	Number of Core Parameters
DQN	33	8983	7
PPO	18	426	9
SAC	121	11937	11

4.2. Optimization Performance Comparison

As a baseline for comparison, a rule-based power-following energy management strategy is adopted. In this strategy, the fuel cell output power primarily follows the instantaneous power demand of the vehicle, while the battery compensates for the power difference to maintain system balance. The control logic does not involve optimization objectives related to efficiency or degradation mitigation and relies on predefined rules only. Under the same driving cycle and initial conditions, this baseline strategy is used to evaluate the reference degradation behavior of the fuel cell system, providing a fair benchmark for assessing the effectiveness of the proposed energy management strategy.

As shown in Figures 13 and 14 and summarized in Table 9, the SAC-based strategy exhibits distinct power allocation patterns compared to DQN, PPO, and power-following strategies. Specifically, SAC allocates a larger proportion of operating points to the medium-to-high efficiency range of the fuel cell, directly enhancing hydrogen utilization efficiency. In contrast, the power trajectories generated by DQN and PPO strategies are smoother with longer steady-state operation times, indicating more conservative power tracking behavior. Quantitative analysis reveals that during the CLTC cycle, the SAC strategy reduces hydrogen consumption by 11.15 g/100 km and 5.54 g/100 km compared to DQN and PPO, respectively. Furthermore, all three reinforcement learning-based strategies achieve significantly lower hydrogen consumption per 100 km than the power-following strategy.

**Figure 13.** SOC under different algorithms in CLTC conditions.**Table 9.** CLTC control results.

Algorithm	Hydrogen Consumption (g)/100 km	Degradation Rate ($\times 10^{-4}\%$)/100 km	Comparison with SAC (Degradation Rate)
Power-following	917.13	9.01	42.84%
DQN	759.33	19.70	73.86%
PPO	753.72	13.68	62.35%
SAC	748.18	5.15	-

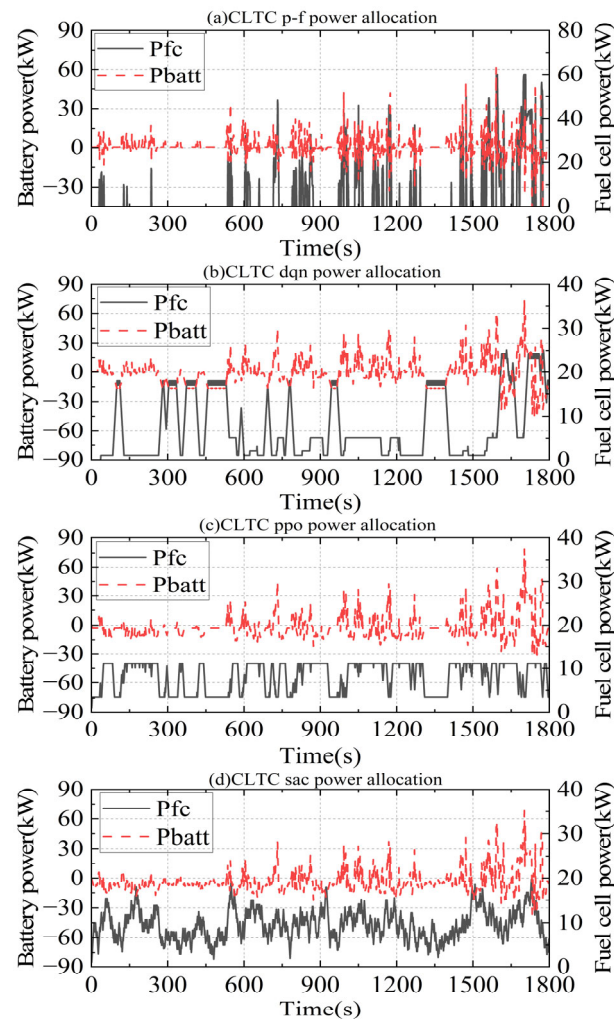


Figure 14. CLTC power distribution.

Despite introducing more frequent load changes, the SAC strategy did not induce acceleration degradation. Conversely, it achieved degradation rates reduced by 73.86% and 62.35% compared to DQN and PPO, respectively, and by 42.84% compared to the power-following strategy. This indicates that degradation depends not only on load smoothness but is significantly influenced by the fuel cell's operating efficiency zone. By prioritizing operation in high-efficiency zones, the SAC-based strategy effectively suppresses degradation mechanisms under enhanced power dynamics, demonstrating its exceptional capability in balancing efficiency and durability.

Compared with DQN and PPO, SAC maintains a higher policy entropy during training due to its maximum-entropy objective, enabling more efficient exploration of the state-action space. This enhanced exploration helps the agent avoid premature convergence to degradation-intensive control patterns, such as frequent load fluctuations or prolonged low-efficiency operation. As training proceeds, the adaptive entropy mechanism gradually reduces stochasticity, allowing the policy to converge to a stable power allocation strategy. This balanced exploration–exploitation process explains why SAC achieves superior degradation suppression while maintaining stable convergence, whereas DQN and PPO tend to converge faster but are more prone to locally optimal solutions with weaker durability awareness.

4.3. Transferability and Robustness Performance

The agent trained on the CLTC cycle was evaluated on the unseen WLTC cycle with an initial SOC of 0.6 to assess generalization capability. As shown in Figures 15 and 16 and

summarized in Table 10, all three reinforcement learning-based strategies exhibit power allocation patterns broadly consistent with those observed during training, indicating stable policy execution under distribution shift. In particular, the SAC-based strategy continues to allocate a higher proportion of operating points within the fuel cell's high-efficiency region, whereas DQN and PPO maintain smoother power trajectories that reflect conservative load-following behavior.

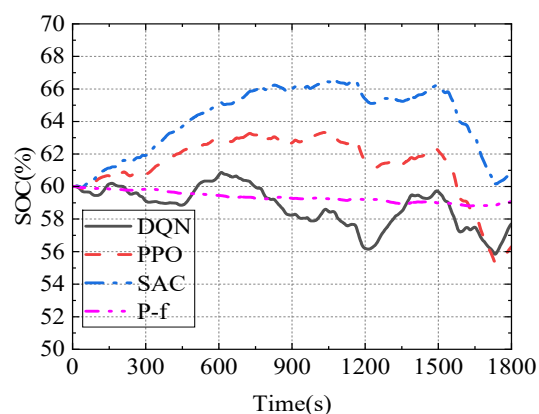


Figure 15. SOC under different algorithms in WLTC conditions.

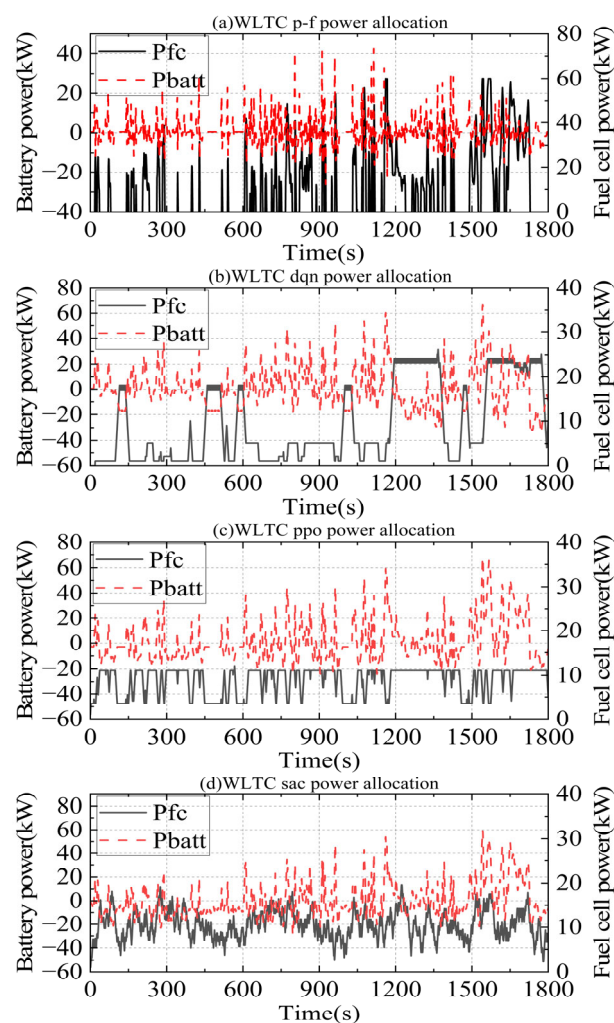


Figure 16. WLTC power distribution.

Table 10. WLTC control results.

Algorithm	Hydrogen Consumption (g)/100 km	Degradation Rate ($\times 10^{-4}\%$)/100 km	Comparison with SAC (Degradation Rate)
Power-following	1105.05	8.06	59.43%
DQN	940.26	17.17	80.96%
PPO	918.26	10.13	67.72%
SAC	911.87	3.27	-

From a performance perspective, SAC preserves its advantage in both economy and durability under the unseen WLTC cycle, achieving reductions in hydrogen consumption of 28.39 g, 6.39 g and 193.18 g per 100 km and degradation reductions of 80.96%, 67.72% and 59.43% relative to DQN, PPO and Power-following, respectively. These results suggest that the degradation-aware reward formulation enables SAC to generalize efficiency-oriented control principles beyond the training cycle. However, the similarity between power distribution patterns in the training and testing cycles also indicates that the learned policy largely reproduces previously acquired behaviors, which may constrain its adaptability to cycle-specific dynamics. This observation highlights a trade-off between policy stability and adaptive responsiveness, suggesting that broader training scenarios or online adaptation mechanisms may be required to further enhance robustness under more diverse driving conditions.

5. Conclusions and Discussion

5.1. Conclusions

This study proposes a degradation-aware reinforcement learning-based energy management strategy for fuel cell hybrid electric vehicles using a Soft Actor–Critic (SAC) framework. By explicitly incorporating hydrogen consumption, SOC regulation, and power fluctuation into the reward function, the proposed method systematically balances economy and durability at the supervisory control level.

Through comprehensive hyperparameter and reward-weight sensitivity analyses, it is shown that SAC performance is primarily governed by network capacity, learning rate, and target entropy, which strongly influence convergence stability and exploration behavior, whereas other parameters exhibit secondary effects. The results further indicate that appropriate degradation-related reward shaping can effectively guide long-term durability without significantly compromising driving range, providing practical insights for DRL-based EMS design.

Comparative evaluations with DQN and PPO demonstrate that, although PPO converges more rapidly and DQN requires fewer tuning parameters, SAC consistently achieves superior trade-offs between hydrogen economy and fuel cell durability under both trained and untrained driving cycles. This advantage is attributed to SAC's entropy-regularized policy, which enables more comprehensive exploration and improved operation within high-efficiency regions of the fuel cell system. The maintained performance under unseen conditions further suggests enhanced robustness and transferability of the learned policy.

Overall, this work highlights the importance of integrating degradation awareness and systematic hyperparameter analysis into reinforcement learning-based energy management strategies. The findings provide methodological guidance for the development of durable and adaptive EMS for fuel cell vehicles and lay the foundation for future extensions toward broader operating scenarios, online adaptation, and experimental validation.

5.2. Discussion

5.2.1. Practical Implications and Applicability

The proposed SAC-based energy management strategy is implemented at the supervisory control level and relies only on measurable system variables commonly available in fuel cell hybrid electric vehicles, which facilitates its practical application without additional hardware requirements. The trained agent learns a continuous control policy through interaction with diverse operating conditions during training, enabling adaptive responses to varying load demands. Although the agent is trained under specific scenarios, the learned policy exhibits a certain degree of generalization within the same system configuration, and its adaptability to broader operating conditions can be further enhanced through retraining or online fine-tuning, which will be investigated in future work.

5.2.2. Limitations and Future Work

This study investigates a degradation-aware reinforcement learning-based energy management strategy mainly through system-level simulations, and several limitations should be acknowledged. First, the degradation model is empirical and adopts constant coefficients without calibration from dynamic aging bench tests, and temperature-, humidity-, and water-management-related effects—such as temperature-accelerated membrane degradation and liquid water-induced reactant transport limitations in the gas diffusion layer under low-load operation—are not explicitly modeled, which constrains the accuracy of absolute lifetime prediction. Second, although temperature-dependent efficiency characteristics are considered, the synergistic optimization between thermal management, efficiency, and durability is not explicitly addressed in the EMS design. Third, the reward function weights are empirically selected without a formal multi-objective optimization framework or systematic sensitivity analysis, and the SAC state space does not include historical load or temperature-related states that may affect fatigue-driven degradation. Finally, generalization performance is evaluated only under an untrained driving cycle without covering standardized or extreme operating conditions such as cold starts or steep gradients. Therefore, the reported results should be interpreted as comparative system-level insights, motivating future work on experimental validation, enhanced multi-physics degradation modeling, thermal–durability co-optimization, and more comprehensive generalization assessment.

Author Contributions: Conceptualization, H.Z. and C.D.; methodology, C.D.; software, H.Z.; validation, H.Z., Y.H. and C.D.; formal analysis, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z., Y.H. and C.D.; project administration, C.D.; funding acquisition, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Foshan Xianhu Laboratory of the Advanced Energy Science and Technology Guangdong Laboratory, grant number XHRD2024-11233100-01.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xian, Y.; Xia, M.; Su, S.; Guo, M.; Chen, F. Research on the market diffusion of fuel cell vehicles in China based on the generalized bass model. *IEEE Trans. Ind. Appl.* **2021**, *58*, 2950–2960. [\[CrossRef\]](#)
2. Navinkumar, T.; Bharatiraja, C. Sustainable hydrogen energy fuel cell electric vehicles: A critical review of system components and innovative development recommendations. *Renew. Sustain. Energy Rev.* **2025**, *215*, 115601. [\[CrossRef\]](#)
3. Li, B.; Wang, C.; Liu, M.; Fan, J.; Yan, J. Transient performance analysis of a solid oxide fuel cell during power regulations with different control strategies based on a 3D dynamic model. *Renew. Energy* **2023**, *218*, 119266. [\[CrossRef\]](#)

4. Burke, A.F.; Zhao, J.; Miller, M.R.; Sinha, A.; Fulton, L.M. Projections of the costs of medium-and heavy-duty battery-electric and fuel cell vehicles (2020–2040) and related economic issues. *Energy Sustain. Dev.* **2023**, *77*, 101343. [\[CrossRef\]](#)
5. Fragiaco, P.; Genovese, M.; Piraino, F.; Corigliano, O.; De Lorenzo, G. Hydrogen-fuel cell hybrid powertrain: Conceptual layouts and current applications. *Machines* **2022**, *10*, 1121. [\[CrossRef\]](#)
6. Ding, Z.; Song, L.; Fang, H.; Yao, Z. A review on sustainability assessment of fuel cell electric vehicles. *Transp. Res. Interdiscip. Perspect.* **2025**, *32*, 101561. [\[CrossRef\]](#)
7. Khalatbarisoltani, A.; Zhou, H.; Tang, X.; Kandidayeni, M.; Boulon, L.; Hu, X. Energy management strategies for fuel cell vehicles: A comprehensive review of the latest progress in modeling, strategies, and future prospects. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 14–32. [\[CrossRef\]](#)
8. Zhang, F.; Wang, L.; Coskun, S.; Pang, H.; Cui, Y.; Xi, J. Energy management strategies for hybrid electric vehicles: Review, classification, comparison, and outlook. *Energies* **2020**, *13*, 3352. [\[CrossRef\]](#)
9. Peng, H.; Li, J.; Thul, A.; Deng, K.; Ünlübayir, C.; Löwenstein, L.; Hameyer, K. A scalable, causal, adaptive rule-based energy management for fuel cell hybrid railway vehicles learned from results of dynamic programming. *Etransportation* **2020**, *4*, 100057. [\[CrossRef\]](#)
10. Deng, L.; Radzi, M.A.M.; Shafie, S.; Hassan, M.K. Optimization of fuel cell switching control based on power following strategy in fuel cell hybrid electrical vehicle. *Int. J. Renew. Energy Dev.* **2025**, *14*, 299–310. [\[CrossRef\]](#)
11. Zhu, Y.; Li, X.; Liu, Q.; Li, S.; Xu, Y. A comprehensive review of energy management strategies for hybrid electric vehicles. *Mech. Sci.* **2022**, *13*, 147–188. [\[CrossRef\]](#)
12. Zhao, X.; Wang, L.; Zhou, Y.; Pan, B.; Wang, R.; Wang, L.; Yan, X. Energy management strategies for fuel cell hybrid electric vehicles: Classification, comparison, and outlook. *Energy Convers. Manag.* **2022**, *270*, 116179. [\[CrossRef\]](#)
13. Wang, Y.; Jiao, X. Dual heuristic dynamic programming based energy management control for hybrid electric vehicles. *Energies* **2022**, *15*, 3235. [\[CrossRef\]](#)
14. Song, K.; Wang, X.; Li, F.; Sorrentino, M.; Zheng, B. Pontryagin’s minimum principle-based real-time energy management strategy for fuel cell hybrid electric vehicle considering both fuel economy and power source durability. *Energy* **2020**, *205*, 118064. [\[CrossRef\]](#)
15. Wang, T.; Li, Q.; Wang, X.; Qiu, Y.; Liu, M.; Meng, X.; Li, J.; Chen, W. An optimized energy management strategy for fuel cell hybrid power system based on maximum efficiency range identification. *J. Power Sources* **2020**, *445*, 227333. [\[CrossRef\]](#)
16. He, H.; Quan, S.; Sun, F.; Wang, Y.-X. Model predictive control with lifetime constraints based energy management strategy for proton exchange membrane fuel cell hybrid power systems. *IEEE Trans. Ind. Electron.* **2020**, *67*, 9012–9023. [\[CrossRef\]](#)
17. Mohammed, A.S.; At Naw, S.M.; Salau, A.O.; Eneh, J.N. Review of optimal sizing and power management strategies for fuel cell/battery/super capacitor hybrid electric vehicles. *Energy Rep.* **2023**, *9*, 2213–2228. [\[CrossRef\]](#)
18. Li, J.; Liu, J.; Yang, Q.; Wang, T.; He, H.; Wang, H.; Sun, F. Reinforcement learning based energy management for fuel cell hybrid electric vehicles: A comprehensive review on decision process reformulation and strategy implementation. *Renew. Sustain. Energy Rev.* **2025**, *213*, 115450. [\[CrossRef\]](#)
19. Wang, H.; He, H.; Bai, Y.; Yue, H. Parameterized deep Q-network based energy management with balanced energy economy and battery life for hybrid electric vehicles. *Appl. Energy* **2022**, *320*, 119270. [\[CrossRef\]](#)
20. Wang, Z.; Zhang, S.; Luo, W.; Xu, S. Deep reinforcement learning with deep-Q-network based energy management for fuel cell hybrid electric truck. *Energy* **2024**, *306*, 132531. [\[CrossRef\]](#)
21. Zheng, M.; Zhang, J.; Zhan, C.; Ren, X.; Lü, S. Proximal policy optimization with reward-based prioritization. *Expert Syst. Appl.* **2025**, *283*, 127659. [\[CrossRef\]](#)
22. Li, B.; Cui, Y.; Xiao, Y.; Fu, S.; Choi, J.; Zheng, C. An improved energy management strategy of fuel cell hybrid vehicles based on proximal policy optimization algorithm. *Energy* **2025**, *317*, 134585. [\[CrossRef\]](#)
23. Lv, X.; Qian, S.; Zhai, X.; Wang, P.; Wu, T. Adaptive energy management strategy for FCHEV based on improved proximal policy optimization in deep reinforcement learning algorithm. *Energy Convers. Manag.* **2024**, *321*, 118977. [\[CrossRef\]](#)
24. Xu, C.; Hayashi, N.; Inuiguchi, M.; Raymond, W.J.K.; Mokhlis, H.; Illias, H.A. Enhanced twin delayed DDPG with prioritized experience replay and Noisy Nets for regional economic dispatch. *Sci. Rep.* **2025**, *15*, 43610. [\[CrossRef\]](#)
25. Yuan, X.; Shang, Z.; Huang, W.; Cui, Y.; Chen, D.; Zhu, M. Effective Reinforcement Learning Control using Conservative Soft Actor-Critic. *arXiv* **2025**, arXiv:2505.03356. [\[CrossRef\]](#)
26. Lu, H.; Tao, F.; Fu, Z.; Sun, H. Battery-degradation-involved energy management strategy based on deep reinforcement learning for fuel cell/battery/ultracapacitor hybrid electric vehicle. *Electr. Power Syst. Res.* **2023**, *220*, 109235. [\[CrossRef\]](#)
27. Wang, Z.; He, H.; Peng, J.; Chen, W.; Wu, C.; Fan, Y.; Zhou, J. A comparative study of deep reinforcement learning based energy management strategy for hybrid electric vehicle. *Energy Convers. Manag.* **2023**, *293*, 117442. [\[CrossRef\]](#)
28. Xu, D.; Cui, Y.; Ye, J.; Cha, S.W.; Li, A.; Zheng, C. A soft actor-critic-based energy management strategy for electric vehicles with hybrid energy storage systems. *J. Power Sources* **2022**, *524*, 231099. [\[CrossRef\]](#)

29. Yang, D.; Wang, L.; Yu, K.; Liang, J. A reinforcement learning-based energy management strategy for fuel cell hybrid vehicle considering real-time velocity prediction. *Energy Convers. Manag.* **2022**, *274*, 116453. [[CrossRef](#)]
30. Zheng, C.; Wang, Y.; Liu, Z.; Sun, T.; Kim, N.; Jeong, J.; Cha, S.W. A hybrid energy storage system for an electric vehicle and its effectiveness validation. *Int. J. Precis. Eng. Manuf.-Green Technol.* **2021**, *8*, 1739–1754. [[CrossRef](#)]
31. Teng, T.; Zhang, X.; Xue, Q.; Zhang, B. Research of proton exchange membrane fuel cell modeling on concentration polarization under variable-temperature operating conditions. *Energies* **2024**, *17*, 730. [[CrossRef](#)]
32. Fan, L.; Xu, K.; Jiang, Z.; Shen, C.; Sun, J.; Wei, Y. Advances of membrane electrode assembly aging research of proton exchange membrane fuel cell under variable load: Degradation mechanism, aging indicators, prediction strategy, and perspectives. *Ionics* **2024**, *30*, 5111–5140. [[CrossRef](#)]
33. Fan, L.; Gao, J.; Lu, Y.; Shen, W.; Zhou, S. Empirical Degradation Models of the Different Indexes of the Proton Exchange Membrane Fuel Cell Based on the Component Degradation. *Energies* **2023**, *16*, 8012. [[CrossRef](#)]
34. Zhang, Y.; Zhu, Q.; Ren, Z. Research on reinforcement learning energy management strategies for fuel cell vehicles considering fuel cell performance degradation. *Electr. Eng.* **2023**, *15*, 63–65. (In Chinese)
35. Yan, S.; Yang, M.; Sun, C.; Xu, S. Liquid water characteristics in the compressed gradient porosity gas diffusion layer of proton exchange membrane fuel cells using the Lattice Boltzmann Method. *Energies* **2023**, *16*, 6010. [[CrossRef](#)]
36. Song, K.; Hou, T.; Jiang, J.; Grigoriev, S.A.; Fan, F.; Qin, J.; Wang, Z.; Sun, C. Thermal management of liquid-cooled proton exchange membrane fuel cell: A review. *J. Power Sources* **2025**, *648*, 237227. [[CrossRef](#)]
37. Tekin, M.; Karamangil, M.İ. Comparative analysis of equivalent circuit battery models for electric vehicle battery management systems. *J. Energy Storage* **2024**, *86*, 111327. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.