applied sciences

MDPI

Article

Optimization Design of Energy Management Strategy for Fuel Cell Hybrid Electric Vehicles

Mengjie Li, Yongbao Liu * and Xing He

Naval University of Engineering, 717 Jiefang Avenue, Wuhan 430000, China; limengjie@wdu.edu.cn (M.L.)
* Correspondence: liuyongbaoly@163.com; Tel.: +86-13986000097

Abstract

To address the limitations of traditional energy management strategies in fuel cell hybrid power systems—specifically their difficulty in simultaneously accommodating dynamic driving condition adaptability, hydrogen fuel economy, and energy storage system stability—this study proposes a power distribution optimization strategy based on Deep Deterministic Policy Gradient (DDPG). The strategy targets a hybrid powertrain architecture dominated by a fuel cell (FC) and assisted by a lithium battery and a supercapacitor. By constructing a multi-dimensional state space that integrates vehicle speed, acceleration, the state of charge (SOC) of the energy storage system, and load power demand, a multi-objective reward function encompassing hydrogen consumption, SOC deviation, system efficiency, and power fluctuation is designed to achieve dynamic power allocation in a continuous action space. Simulation studies are conducted under three typical driving cycles—WLTP, CLTC-P, and UDDS—with comparative evaluations against the conventional Equivalent Consumption Minimization Strategy (ECMS) and Deep Q-Network (DQN)-based strategies. The results demonstrate that the DDPG-based strategy reduces hydrogen consumption to 607.1 g/100 km, 580.2 g/100 km, and 560.0 g/100 km under the three driving cycles, respectively, achieving a maximum reduction of 28% compared with ECMS. The average system efficiency increases to 64–66%, representing an improvement of 38.9%, while the operating proportion of the fuel cell within the high-efficiency region (40–80% load) increases by 15%. In addition, the strategy exploits the high-frequency response capability of the supercapacitor to smooth instantaneous power fluctuations, effectively reducing the inefficient start–stop events of the fuel cell. Although the SOC fluctuation range of the lithium battery increases by 32.5% compared with ECMS, a dynamic balance between energy efficiency and battery lifespan can be achieved through optimized weighting of the SOC deviation penalty term in the reward function. Overall, this study provides a solution with both theoretical significance and engineering feasibility for global energy optimization of fuel cell–energy storage systems under complex driving conditions.

Keywords: fuel cell hybrid electric vehicle; energy management strategy; deep deterministic policy gradient (DDPG); power distribution optimization; hydrogen consumption reduction

Check for updates

Academic Editor: Jun Cong Ge

Received: 12 March 2026
Revised: 29 April 2026
Accepted: 30 April 2026
Published: 4 May 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) license.

1. Introduction

With the global energy transition and the advancement of the "dual carbon" goals, fuel cell hybrid electric vehicles (FCHEVs) have attracted significant attention as a key technological pathway toward decarbonization in the transportation sector, owing to their core advantages of near-zero emissions, long driving range, and fast refuelling capabilities. It should be noted that hydrogen, as the primary fuel for FCHEVs, is not freely available in

nature in large quantities and must be produced through dedicated processes. Renewable energy sources are widely recognized as effective alternatives to fossil fuels. However, due to the inherent intermittency and fluctuation of natural resources such as wind and solar energy, their stability in energy supply and large-scale deployment have been limited [1]. Hydrogen energy holds significant potential in the clean energy transition and serves as an effective pathway for achieving large-scale deep decarbonization. Specifically, electrolysis of water powered by curtailed wind and solar electricity for green hydrogen production is not only a primary method for obtaining hydrogen energy, but also contributes to addressing the intermittency and variability of renewable energy sources. This synergy between green hydrogen production and renewable energy curtailment absorption has created a favorable opportunity for the rapid development of the fuel cell vehicle industry, providing a sustainable and scalable hydrogen supply chain for FCHEVs. The energy management strategy (EMS), known as the "brain" of the FCHEV, aims to achieve optimal power distribution among the fuel cell (FC), lithium-ion battery, and supercapacitor within the hybrid energy storage system (ESS) [2]. Its primary objectives are to balance hydrogen economy, fuel cell durability, and the stability of the state of charge (SOC). However, the frequent start–stop events, acceleration, deceleration, and sudden slope variations during real-world driving result in highly transient power demands, posing significant challenges for dynamic energy optimization [3].

Conventional energy management strategies often struggle to adapt to such dynamic and complex conditions. Rule-based control strategies rely on predefined logic, offering real-time response but lacking global optimization capabilities, thus failing to address transient load fluctuations effectively. The equivalent consumption minimization strategy (ECMS) transforms electric energy loss into equivalent hydrogen consumption, achieving near-global optimization. However, fixed equivalence factors may lead to fuel cell overloading or deep battery discharges, accelerating component degradation. Model predictive control (MPC) depends on accurate dynamic models [4–8]; yet, due to nonlinear fuel cell behavior and energy storage aging, model mismatches often degrade control precision. PI and state machine controls are simple and robust but can only achieve local optimization, falling short in addressing the multi-objective nature of EMS design. These limitations hinder conventional methods from achieving coordinated optimization among economy, durability, and system stability under dynamic operating conditions.

In recent years, deep reinforcement learning (DRL) has emerged as a promising solution for complex EMS problems, owing to its model-free adaptability [9–13], autonomous learning, and dynamic decision-making capabilities. The deep Q-network (DQN), a classical DRL algorithm, optimizes power distribution within a discrete action space. However, the discretization process can introduce quantization errors, reducing control accuracy. In contrast, the deep deterministic policy gradient (DDPG) algorithm employs an actor–critic framework that outputs continuous control actions, aligning well with the continuous power requirements of hybrid powertrains and effectively avoiding discretization errors. Moreover, DDPG leverages experience replay and target network soft-update mechanisms to break data correlations and enhance training stability, making it suitable for long-term optimization tasks under dynamic driving conditions.

To address the shortcomings of traditional methods, this study proposes a DDPG-based optimal power allocation strategy for the fuel cell–energy storage system. A multidimensional state space incorporating vehicle speed, acceleration, load power demand, and SOC is first constructed to characterize the system's operational state comprehensively. Then, a multi-objective reward function integrating hydrogen consumption, SOC deviation, system efficiency, and power fluctuation is designed to guide the agent toward learning globally optimal control policies. Finally, the proposed strategy is validated under three

standard driving cycles—WLTP, CLTC-P, and UDDS—and benchmarked against ECMS and DQN strategies. The goal is to leverage DDPG's dynamic optimization capabilities to achieve hydrogen consumption minimization, system efficiency maximization, and SOC stability simultaneously, thereby providing theoretical insights and engineering references for the practical application of FCHEV energy management technologies.

2. Fuel Cell-Based Hybrid Energy Control Structure

In this section, a simulation framework of the fuel cell hybrid electric vehicle (FCHEV) energy management system is developed to illustrate the power flow and energy distribution mechanisms among multiple energy sources. This structure provides a foundation for analyzing the energy output, consumption, and power distribution characteristics of fuel cells and energy storage devices, as well as for evaluating the impact of various control strategies and algorithms on system performance. By designing a fuel cell–based hybrid energy control structure, the optimal energy management strategy and parameter configuration can be determined, thereby improving overall system efficiency and enhancing vehicle performance.

Figure 1 shows the energy control architecture for the hybrid system. Several commonly used control strategies and algorithms are incorporated within this framework, including PI control, Equivalent Consumption Minimization Strategy (ECMS), Energy Estimation and Management System (EEMS), frequency decoupling and state machine control, and pure state machine control. A performance comparison among these control strategies is summarized in Table 1.

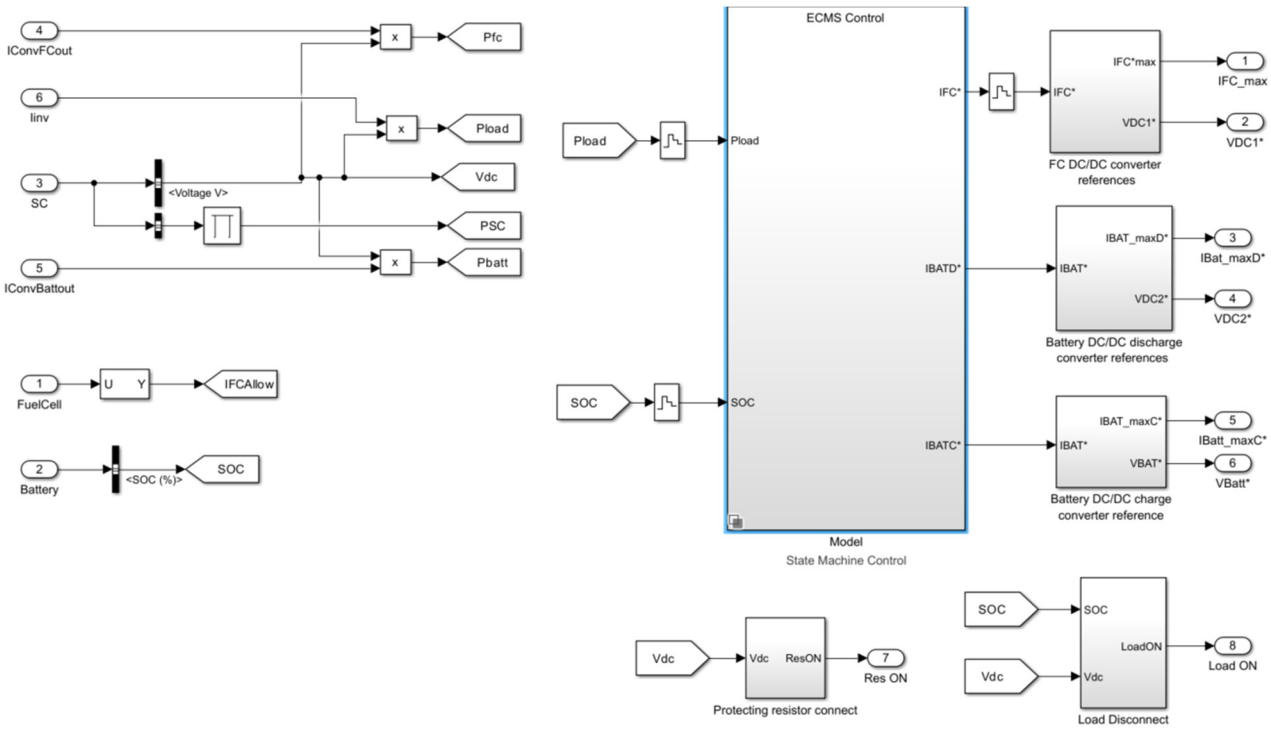

Figure 1. Energy control structure for fuel cell hybrid systems.

In dynamic driving scenarios, it is crucial to achieve a global optimization of energy distribution between the fuel cell and the energy storage batteries [14–16]. Compared with the local regulation limitation of traditional PI control, the Equivalent Consumption Minimization Strategy (ECMS) introduces an equivalence factor to unify hydrogen consumption and electrical energy consumption into a common metric. This allows for a dynamic balance between instantaneous power demand and long-term energy efficiency, thereby

significantly reducing energy waste while ensuring the vehicle's power responsiveness. Although its computational complexity is higher than that of basic control methods, ECMS is more feasible for engineering implementation compared with algorithms such as EEMS, which rely heavily on high-precision modeling or real-time big data analysis. In addition, this strategy has been widely applied and validated in hybrid power systems. Its algorithmic framework can adapt to different vehicle types and operating conditions through adaptive parameter tuning, combining theoretical rigor with engineering adaptability, and ultimately achieving the most rational trade-off among energy efficiency, dynamic performance, and implementation cost.

Table 1. Comparison of control strategies for energy management.

Control Strategy	Control Objective	Complexity	Real-Time	Optimization Capability	Applicable Scene	Merit	Defect
PI control	Stable output power to meet power requirements	Low	High	Local optimization	steady-state or low dynamic demand scenarios	Simple, reliable, and flexible parameter adjustment	Insufficient dynamic performance for global optimization
ECMS	Minimum equivalent energy consumption	Medium to high	Middle	Overall optimization	dynamic energy allocation requirement scenario	Clear objectives and high efficiency in energy allocation	Depends on model accuracy and has high computational complexity
EEMS	Maximization of energy efficiency	High	Low	Overall optimization	Multi-energy collaborative management scenario	Data-driven, adaptable to complex working conditions	High computational resource requirements and poor real-time performance
Frequency Decoupling and State Machine Control	Power Decoupling and Multi-system Cooperative Control	High	Medium to high	Local + collaborative optimization	Multi-energy independent regulation scenario	Decoupling strategy is flexible and adaptable to dynamic working conditions	Modeling is complex and debugging is challenging
state machine control	Switch system status to match requirements	Middle	High	State-level optimization	Discrete operating condition switching scenario	Logical and responsive	Relying on experience-based design, global optimization capability is limited

For a hybrid power system primarily driven by a fuel cell (FC) and supplemented by a lithium-ion battery (Batt) and a supercapacitor (SC) [17], the core objective of ECMS is to dynamically optimize energy distribution by converting both the hydrogen consumption of the fuel cell and the electrical energy loss of the storage system into equivalent hydrogen consumption, thereby achieving global energy efficiency optimization. Its principle is based on real-time power demand and the state of charge (SOC) of the energy storage system, dynamically adjusting the weighting factors of each energy source through the equivalence factor. The specific derivation is as follows:

The instantaneous power balance of the hybrid system must satisfy the following equation:

$$P_{\text{req}}(t) = P_{\text{fc}}(t) + P_{\text{batt}}(t) + P_{\text{sc}}(t) \quad (1)$$

where P_{req} denotes the vehicle's required power, and P_{fc}, P_{batt}, P_{sc} represent the output power (positive for discharge, negative for charge) of the fuel cell, lithium battery, and supercapacitor, respectively. The ECMS objective function is defined as minimizing the equivalent hydrogen consumption rate.

$$\min\left[\dot{m}_{fc} + s_{batt} \cdot \frac{P_{batt}}{\eta_{batt}} + s_{sc} \cdot \frac{P_{sc}}{\eta_{sc}}\right] \tag{2}$$

In this formula, $\dot{m}_{fc}$ denotes the hydrogen consumption rate of the fuel cell (correlated with power P_{fc}), while s_{batt}, s_{sc} are equivalent factors used to convert the charge–discharge power of lithium batteries and supercapacitors into equivalent hydrogen consumption. The charge–discharge efficiencies of these components are represented [18–22] by η_{batt}, η_{sc}, respectively.

The equivalence factors s_{batt} and s_{sc} are critical parameters in the ECMS formulation, as their values directly determine the trade-off between hydrogen consumption and electrical energy utilization. Improper calibration of these factors may cause the optimization objective to deviate, leading to either excessive depletion or over-conservation of energy storage devices.

In this study, the equivalence factors are determined through an adaptive approach based on real-time SOC feedback. The nominal equivalence factor for the lithium battery is initially estimated from the average fuel cell system efficiency and the battery charge–discharge efficiency:

$$S_{batt,0} = \frac{\overline{\eta}_{fc}}{\eta_{batt,chg} \cdot \eta_{batt,dis}} \tag{3}$$

where $\overline{\eta}_{fc}$ denotes the average fuel cell system efficiency, and $\eta_{batt,chg}$, $\eta_{batt,dis}$ represent the charging and discharging efficiencies of the lithium battery, respectively. The equivalence factor is then dynamically adjusted based on the SOC deviation from the target value:

$$s_{batt} = s_{batt,0} + k_p \cdot (SOC_{ref} - SOC_{batt}) + k_i \cdot \int_0^t (SOC_{ref} - SOC_{batt}) d\tau \tag{4}$$

where SOC_{ref} s the target state of charge, k_p and k_i are the proportional and integral control coefficients designed to suppress SOC deviation. This PI-based adaptive mechanism ensures that when the battery SOC drops below the target, the equivalence factor increases to discourage further discharge; conversely, when SOC exceeds the target, the factor decreases to encourage battery utilization.

For the supercapacitor, given its characteristics of rapid response and extended cycle life, the equivalence factor is simplified as

$$s_{sc} = \alpha \cdot \frac{\overline{\eta}_{fc}}{\eta_{sc}} \tag{5}$$

where α is the dynamic compensation coefficient that focuses on optimizing instantaneous power distribution rather than long-term energy balance.

The hydrogen consumption rate can be calculated using the fuel cell efficiency model:

$$\dot{m}_{fc} = \frac{P_{fc}}{\eta_{fc} \cdot LHV_{H2}} \tag{6}$$

In the formula, η_{fc} denotes the fuel cell efficiency (nonlinearly varying with power), while LHV_{H2} represents the low heating value of hydrogen (approximately 120 MJ/kg). The efficiency curve is typically derived from experimental data fitting [23], for instance:

$$\eta_{fc} = a \cdot P_{fc}^2 + b \cdot P_{fc} + c \quad (7)$$

The parameters a, b, and c in Equation (4) are identified using a nonlinear least squares fitting method based on the polarization curve data provided by the fuel cell manufacturer. It should be noted that the accuracy of fuel cell model parameters directly affects the state space definition and reward function calculation in the reinforcement learning algorithm. To enhance the accuracy of PEMFC models under dynamic conditions, recent work [24] proposed an advanced parameter estimation method based on metaheuristic optimization algorithms, which provides a promising direction for further improving model fidelity.

The equivalent factors s_{batt}, s_{sc} must be updated in real time based on the SOC state of the energy storage system to ensure proper energy storage. The equivalent factor design for lithium batteries is as follows:

$$s_{batt}(t) = s_{base} + k_p(SOC_{ref} - SOC_{batt}(t)) + k_{i\int(SOC_{ref} - SOC_{batt}(\tau))d} \quad (8)$$

In the formula, SOC_{ref} denotes the target state of charge (SOC) for lithium batteries, while k_p, k_i are PI control coefficients designed to suppress SOC deviation. Supercapacitors [25], characterized by rapid response and extended cycle life, can have their equivalent factor simplified as

$$s_{sc}(t) = s_{base} + \alpha(SOC_{ref,sc} - SOC_{sc}(t)) \quad (9)$$

In the formula, α is the dynamic compensation coefficient, which focuses on optimizing the instantaneous power distribution [26].

The ECMS solution must satisfy both power balance and equipment limit constraints:

$$\begin{cases} P_{fc,min} \le P_{fc} \le P_{fc,max} \\ P_{batt,min} \le P_{batt} \le P_{batt,max} \\ P_{sc,min} \le P_{sc} \le P_{sc,max} \\ SOC_{batt,min} \le SOC_{batt} \le SOC_{batt,max} \\ SOC_{sc,min} \le SOC_{sc} \le SOC_{sc,max} \end{cases} \quad (10)$$

By applying the Lagrange multiplier method to find the extremum of the objective function and combining it with the power balance equation $P_{batt} + P_{sc} = P_{req} - P_{fc}$, we derive the optimal conditions:

$$\frac{dm_{fc}}{dP_{fc}} = -\left(\frac{s_{batt}}{\eta_{batt}} + \frac{s_{sc}}{\eta_{sc}}\right) \quad (11)$$

This condition indicates that the marginal hydrogen consumption of the fuel cell must equal the equivalent marginal cost of the energy storage system. In practical applications, this equation is solved by consulting tables or numerical iteration to obtain the optimal power P_{fc}^*, with the remaining power $P_{req} - P_{fc}^*$ allocated to lithium batteries and supercapacitors according to priority [27–30].

ECMS achieves cost parity between hydrogen and electricity through equivalent factors, dynamically adjusting energy allocation weights via SOC feedback. This mechanism not only maintains fuel cells' optimal operating range but also utilizes supercapacitors to smooth power fluctuations and extend lithium battery lifespan. Its global optimization capability outperforms conventional rule-based strategies under dynamic conditions while avoiding computational complexity issues typical of dynamic programming algorithms, making it the ideal energy management solution for fuel cell hybrid systems [31].

3. Design of Energy Management Strategy Based on Reinforcement Learning

3.1. Definition of State Space and Action Space

Reinforcement learning primarily consists of five key components: the agent, environment, state, actions, and reward function. The core mechanism involves the agent iteratively testing and refining its actions through environmental interactions, while the environment updates its state based on feedback rewards to continuously improve the agent's strategy until convergence. The energy management strategy framework in reinforcement learning is illustrated in Figure 2.. Acting as the agent, the energy management strategy receives the current state s_t and reward r_t, then generates the action a_t to influence the environment. This process updates the environmental state according to the established fuel cell vehicle model, repeating the cycle until convergence is achieved.

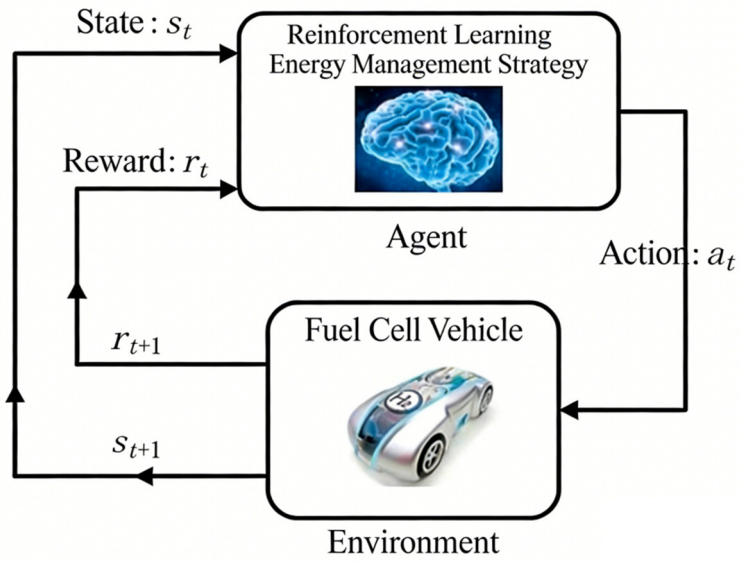

Figure 2. Reinforcement learning framework for energy management.

The objective of reinforcement learning algorithms is to maximize the expected cumulative reward that the agent receives from the environment.

$$V^*(s) = \mathrm{E}\left(\sum_{t=0}^{\infty} \gamma^t r_t\right) \tag{12}$$

In this formula, r denotes the discount rate, E represents the expectation, r_r is the instantaneous reward at each time step, and $V^*(s)$ satisfies the Bellman equation:

$$V^*(s) = \mathrm{r}(s,a) + \gamma \sum_{s=S}^{+\infty} P_{sa,s'} \mathrm{V}^*(s') \tag{13}$$

In this formula, $P_{sa,s'}$ denotes the state transition probability, indicating that the value function corresponding to a state consists of the current instant reward and the discounted future cumulative reward.

In reinforcement learning, the state space and action space are the core components that define the interaction between the agent and the environment. They directly determine the optimization capability and practical performance of the reinforcement learning algorithm. In the context of energy management for fuel cell hybrid systems, the state space describes the current operating condition of the system, while the action space defines the power distribution operations that the agent can execute at each state. A well-designed state

and action space provides the agent with comprehensive environmental information and ensures the feasibility and effectiveness of its decision-making process.

The state space of a fuel cell hybrid power system typically includes key variables such as the state of charge (SOC) of the energy storage device, load power demand, fuel cell output power, and energy storage system output power. The SOC reflects the current energy level of the storage device and serves as an essential basis for optimizing power distribution. The load power demand represents the vehicle's instantaneous power requirement, providing the agent with direct input for decision-making. The output power of the fuel cell and the energy storage system is used to evaluate their respective operating conditions and energy distribution performance. Additionally, system efficiency can also be included as a state variable to further optimize fuel economy. By incorporating these variables, the agent gains a comprehensive understanding of the system's operational status, thereby supporting more informed and adaptive control decisions.

The action space defines the set of power distribution operations that the agent can perform under each state. In a fuel cell hybrid system, the actions typically correspond to the power allocation between the fuel cell and the energy storage device, subject to the constraint that the total output power must equal the load demand. The design of the action space must account for the physical limitations of the system, such as the power bounds of the fuel cell and energy storage components and the permissible SOC range, to ensure that all actions are feasible. For continuous action spaces, algorithms such as the Deep Deterministic Policy Gradient (DDPG) can be employed to output continuous control actions directly; for discrete action spaces, algorithms such as Deep Q-Network (DQN) can be used to select discrete control actions.

Table 2 presents a typical design of the state and action spaces for a fuel cell hybrid power system. By carefully selecting appropriate state variables and defining the action space properly, the overall performance and optimization capability of the reinforcement learning algorithm can be significantly improved.

Table 2. State and action space definitions.

Type	Variable	Describe	Range Constraint
state space	SOC	State of Charge of Energy Storage Device	$0 \leq SOC \leq 100\%$
	load power requirement P_{load}	Current power demand of the vehicle	$P_{load} \geq 0$
	fuel cell output power P_{fc}	Current output power of fuel cell	$0 \leq P_{fc} \leq P_{fc,max}$
	output power of energy storage device P_{ess}	Current output power of energy storage device	$P_{ess,min} \leq P_{ess} \leq P_{ess,max}$,
	system efficiency η	overall energy conversion efficiency of the system	$0 \leq \eta \leq 1$
motion space	fuel cell output power P_{fc}	Fuel Cell Output Power Adjusted by Agent	$0 \leq P_{fc} \leq P_{fc,max}$
	output power of energy storage device P_{ess}	Output Power of Energy Storage Device Adjusted by Agent	$P_{ess,min} \leq P_{ess} \leq P_{ess,max}$

In practical applications, the design of state and action spaces requires comprehensive consideration of system complexity, real-time requirements, and optimization objectives. For instance, by incorporating State of Charge (SOC), load power demands, and fuel cell output power as state variables, the agent can dynamically adjust power allocation strategies based on current conditions to optimize fuel efficiency and SOC maintenance. Furthermore, the design of action spaces must account for physical constraints such as power limitations in fuel cells and energy storage systems to ensure feasible actions.

3.2. Design of Reward Function and Experience Replay

In reinforcement learning, the reward function and experience replay are two key mechanisms that enable the agent to learn an optimal control policy. The reward function provides the direction and criteria for optimization, guiding the agent toward desired behavior, while the experience replay mechanism improves training stability and efficiency by storing and randomly sampling past interaction data. In the context of energy management for fuel cell hybrid systems, a well-designed reward function and an optimized experience replay strategy can significantly enhance both the learning performance of the agent and the overall system efficiency.

The design of the reward function must take into account multiple optimization objectives, including fuel economy, state-of-charge (SOC) maintenance, system efficiency, and dynamic response performance. By introducing several penalty terms with appropriate weighting coefficients, the reward function can effectively balance these often competing objectives. In a fuel cell hybrid power system, the reward function typically consists of the following components:

1. Fuel consumption penalty term—This term is designed to minimize the hydrogen consumption of the fuel cell and can be expressed as

$$R_{\text{fuel}}(t) = -\alpha \cdot \dot{m}_{\text{fc}}(t) \tag{14}$$

In this formula, $\dot{m}_{\text{fc}}(t)$ denotes the instantaneous fuel consumption rate of the fuel cell, while α represents the weighting coefficient.

2. SOC deviation penalty term: Maintains SOC within the target range, expressed as

$$R_{\text{SOC}}(t) = -\beta \cdot (\text{SOC}(t) - \text{SOC}_{\text{ref}})^2 \tag{15}$$

In this formula, $\text{SOC}(t)$ denotes the current SOC value, SOC_{ref} represents the target SOC value, and β is the weight coefficient.

3. System efficiency penalty term: Used to optimize the energy conversion efficiency of the system, its expression is

$$R_{\text{efficiency}}(t) = -\gamma \cdot (1 - \eta(t)) \tag{16}$$

In this formula, $\eta(t)$ denotes the system's instantaneous efficiency, while γ represents the weighting coefficient.

4. Dynamic response penalty term: Used to optimize the system's dynamic response performance, expressed as

$$R_{\text{dynamic}}(t) = -\delta \cdot (P_{\text{load}}(t) - P_{\text{fc}}(t) - P_{\text{ess}}(t))^2 \tag{17}$$

In this formula, $P_{\text{load}}(t)$ denotes the load power demand, while $P_{\text{fc}}(t)$, $P_{\text{ess}}(t)$ represent the output power of the fuel cell and energy storage system respectively, with δ being the weighting coefficient.

In summary, the reward function can be expressed as

$$\text{R}(t) = R_{\text{fuel}}(t) + R_{\text{SOC}}(t) + R_{\text{efficiency}}(t) + R_{\text{dynamic}}(t) \tag{18}$$

Table 3 presents the weight coefficients of penalty terms and their design rationale. By appropriately setting these coefficients, the relationships among multiple optimization objectives can be balanced, ensuring the effectiveness of the reward function in practical applications.

Table 3. Reward function weight coefficients.

Penalty Term	Weight Coefficient	Design Basis
fuel consumption penalty	α = 1.0	Fuel economy is the core goal of energy management, so it is given a high weight coefficient to optimize fuel consumption.
SOC bias penalty term	β = 0.5	SOC maintenance is an important factor affecting the life of energy storage device and system stability, and a medium weight coefficient is assigned to balance its importance.
system efficiency penalty term	γ = 0.2	System efficiency directly reflects the overall energy conversion quality of the powertrain. While it is partially correlated with fuel consumption, an independent moderate weight coefficient is assigned to ensure explicit optimization of energy conversion performance without causing conflicting gradients with the fuel consumption term.
dynamic response penalty term	δ = 0.1	Dynamic response performance has an impact on the stability of the system, but its importance is lower than that of fuel economy and SOC maintenance.

Experience replay is a pivotal technique in deep reinforcement learning, designed to store agent–environment interaction data and randomly sample during training to eliminate temporal correlations between data, thereby enhancing training stability and efficiency. In fuel cell hybrid system energy management, this mechanism effectively addresses data correlation issues and low sample efficiency encountered by reinforcement learning algorithms during training, accelerating agent learning and improving optimization outcomes.

The core concept of experience replay involves storing the agent's interaction data with the environment (including state, actions, rewards, and next state) in a replay buffer, with random sampling during training to update the network. The mathematical representation of the replay buffer is

$$\mathrm{D} = \{(s_t,\ a_t,\ r_t,\ s_{t+1})\} \tag{19}$$

Here, D denotes the experience pool, s_t represents the state at time t, a_t indicates the action taken at time t, r_t is the reward received at time t, and s_{t+1} denotes the state at time t + 1.

During training, a minibatch is randomly sampled from the experience pool to update the neural network's parameters. The mathematical expression for experience replay is

$$\mathrm{B} = \left\{(s_i,\ a_i,\ r_i,\ s_{i+1})\right\}_{i=1}^{N} \tag{20}$$

Here, B denotes the minibatch size, and N represents the sample data dimension.

The reward function and experience replay mechanism work synergistically in reinforcement learning. The reward function provides the agent with clear learning objectives and optimization directions, while the experience replay mechanism enhances training stability and efficiency by storing and randomly sampling historical data. In fuel cell hybrid system energy management, through proper design of the reward function and optimization of the experience replay mechanism, the agent can more effectively learn optimal power distribution strategies, thereby significantly improving the system's overall performance.

3.3. Training and Convergence Analyses

In reinforcement learning, training and convergence analyses are critical processes to ensure that the agent effectively learns an optimal control policy. During training, the agent continuously interacts with the environment to iteratively optimize the parameters of its policy network, while convergence analysis evaluates whether the learning process has achieved the desired performance. In the context of energy management for fuel cell

hybrid systems, both theoretical validation and experimental verification are essential to confirm the effectiveness of the learned strategy. By combining real-world or simulated driving data, the training process and the optimization capability of the agent can be demonstrated in a more intuitive and quantitative manner. Throughout training, the agent interacts with the environment repeatedly, adjusting the parameters of the policy and value networks to maximize the expected cumulative reward. It should be clarified that the DDPG algorithm is designed to maximize the cumulative reward. In the reward function design adopted in this study, hydrogen consumption and SOC deviation are formulated as negative penalty terms, while system efficiency is formulated as a positive reward term. Therefore, maximizing the total reward is mathematically equivalent to minimizing hydrogen consumption and SOC deviation while simultaneously improving system efficiency. This iterative optimization process allows the agent to progressively improve its power distribution decisions under varying operating conditions.

To comprehensively evaluate the generalization ability and robustness of the Deep Reinforcement Learning-based Energy Management Strategy (DRL-EMS), three internationally recognized driving cycles—WLTP, CLTC-P, and UDDS—were selected as the training and validation scenarios. The networks were initialized with identical parameters, and the agent was trained for 2000 episodes under each condition. The convergence behavior of the cumulative reward across different driving cycles is illustrated in Figure 3. The horizontal axis represents the number of training episodes (ranging from 0 to 2000, with tick marks at intervals of 400), and the vertical axis represents the average reward value per episode.

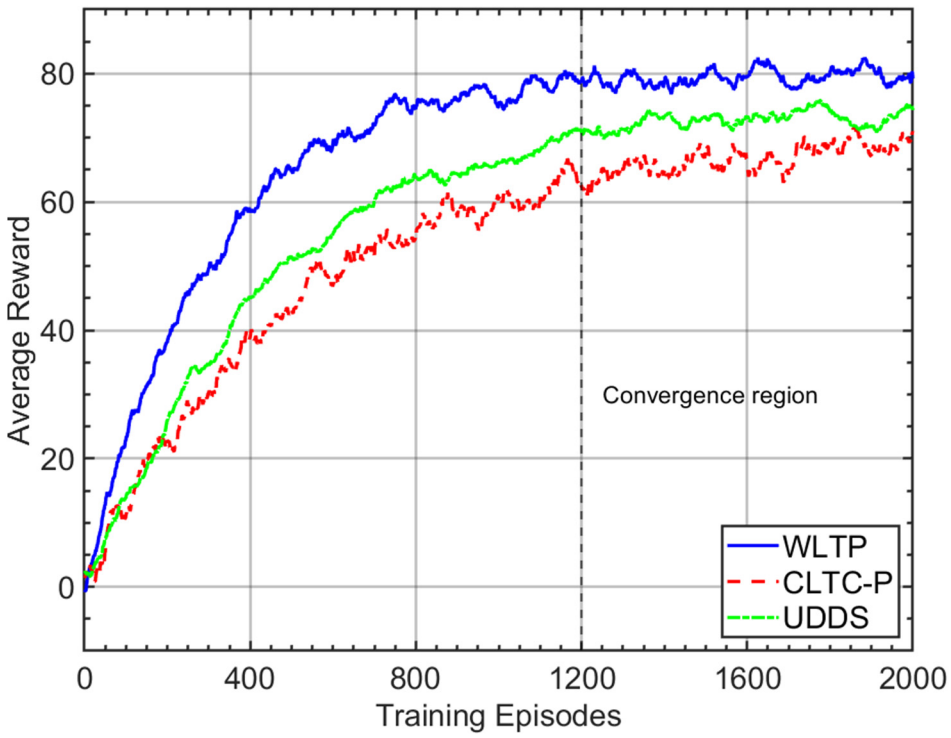

Figure 3. Training reward curves for different driving cycles.

The results demonstrate that the WLTP test cycle achieves significantly faster convergence, with reward values stabilizing after approximately 800 episodes. Both CLTC-P and UDDS exhibit similar convergence rates, requiring around 1200 episodes to reach steady-state. The upward trend of all three reward curves confirms that the agent progressively learns to reduce hydrogen consumption and maintain SOC stability through the reward maximization mechanism. Although these standards originate from different regions, they share urban driving conditions. Their comparable low-speed segment proportions (45%

for CLTC-P versus 50% for UDDS) result in high local power demand similarity, requiring agents to repeatedly refine strategies to accommodate subtle variations.

Furthermore, the training process incorporates a learning rate adaptive decay mechanism (initial learning rate 0.001, with a 10% reduction every 500 episodes) to prevent gradient oscillation in later stages and ensure convergence stability. Notably, the dynamic load characteristics of WLTP enable the agent to rapidly capture power distribution patterns in early stages, whereas CLTC-P and UDDS require longer periods for fine-tuning the energy storage system's charge–discharge thresholds due to frequent start–stop cycles.

The complete hyperparameter configuration of the DDPG algorithm adopted in this study is summarized in Table 4. The Actor network and Critic network each comprise two fully connected hidden layers with 256 and 128 neurons, respectively, employing ReLU activation functions. The learning rates for the Actor and Critic networks are set to 1×10^{-4} and 1×10^{-3}, respectively, following the common practice of assigning a higher learning rate to the Critic to accelerate value estimation convergence. The discount factor γ is set to 0.99 to emphasize long-term cumulative rewards. The soft update coefficient τ for the target networks is set to 0.005 to ensure training stability. The experience replay buffer capacity is set to 100,000 transitions, and the mini-batch size is 64. For exploration, Ornstein–Uhlenbeck (OU) process noise is applied to the Actor's output, with the noise scale linearly decaying from 0.3 to 0.05 over training.

Table 4. Hyperparameter configuration for the DRL-based energy management strategy.

Hyperparameter	Value
Actor hidden layers	256, 128
Critic hidden layers	256, 128
Actor learning rate	1×10^{-4}
Critic learning rate	1×10^{-3}
Discount factor γ	0.99
Soft update coefficient τ	0.005
Replay buffer size	100,000
Mini-batch size	64
OU noise (initial $\rightarrow$ final)	0.3 $\rightarrow$ 0.05
Training episodes	2000

These hyperparameters were initially selected based on widely adopted values in the deep reinforcement learning literature. A systematic grid search was then conducted over representative ranges—learning rate $\in \{5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, batch size $\in \{32, 64, 128\}$, and $\tau \in \{0.001, 0.005, 0.01\}$—using the WLTP cycle as the tuning benchmark. The final configuration was selected based on convergence speed and steady-state hydrogen consumption. Its robustness was confirmed by consistent performance under CLTC-P and UDDS without re-tuning.

To quantify the optimization effects of DRL-EMS, closed-loop tests were conducted on the trained agents under three operational conditions. Key performance metrics were compared at 0 (initial policy), 500,1000, and 2000 episodes, with results presented in Tables 5–7.

Table 5. WLTP training iteration performance.

Iterations	Fuel Consumption (g)	Average System Efficiency (%)	Battery SOC Fluctuation Range (%)
0	150.0	47.5	20.5
500	140.3	55.6	27.8
1000	135.6	60.7	30.3
2000	127.1	66.0	32.5

Table 6. CLTC-P training iteration performance.

Iterations	Fuel Consumption (g)	Average System Efficiency (%)	Battery SOC Fluctuation Range (%)
0	110.5	45.3	12.5
500	90.5	53.6	15.3
1000	81.2	60.1	17.6
2000	75.6	64.2	22.8

Table 7. UDDS training iteration performance.

Iterations	Fuel Consumption (g)	Average System Efficiency (%)	Battery SOC Fluctuation Range (%)
0	86.4	48.3	13.2
500	79.8	57.6	14.7
1000	66.7	63.8	18.6
2000	60.0	65.0	20.6

As shown in Table 5, hydrogen consumption decreased from 150.0 g to 127.1 g (a 15.3% reduction) with increasing iterations (0–2000), while the average system efficiency improved significantly (47.5% $\rightarrow$ 66.0%, a 38.9% increase). This optimization occurs because the DDPG agent progressively learns to operate the fuel cell within its high-efficiency region (typically 40–80% of rated power) rather than allowing frequent fluctuations across the entire power range. By maintaining the fuel cell in this optimal operating zone, the agent minimizes hydrogen consumption per unit of energy output. The improved system efficiency is attributed to two key mechanisms: first, the agent learns to leverage the supercapacitor's high power density and fast response characteristics to handle transient load peaks, reducing the fuel cell's burden during rapid power changes; second, the agent optimizes the power split ratio to minimize conversion losses in both the fuel cell stack and the DC/DC converters. However, the battery State of Charge (SOC) fluctuation range expanded (20.5% $\rightarrow$ 32.5%), which is an intentional trade-off rather than a deficiency. The increased SOC variation indicates that the DDPG strategy actively utilizes the energy storage system's buffering capacity to absorb load fluctuations and enable the fuel cell to maintain steady-state operation. This approach prioritizes immediate fuel economy over minimizing battery cycling, as the reward function assigns higher weight to hydrogen consumption reduction ($\alpha = 1.0$) compared to SOC deviation penalty ($\beta = 0.5$). While this strategy enhances short-term energy efficiency, the long-term impact on battery degradation requires further experimental validation under extended driving cycles.

As shown in Table 6, fuel consumption decreased significantly from 110.5 g to 75.6 g (a 31.6% reduction) with 2000 iterations, while the average system efficiency increased from 45.3% to 64.2% (a 41.7% improvement). The substantial performance gain under the CLTC-P cycle is attributed to the cycle's characteristic frequent acceleration–deceleration patterns, which create opportunities for the DDPG agent to exploit the complementary advantages of different power sources. During optimization, the battery SOC fluctuation range expanded from 12.5% to 22.8% (an 82.4% increase), indicating that the agent learned to implement aggressive energy buffering strategies. Specifically, the supercapacitor handles high-frequency power fluctuations (sub-second transients during rapid acceleration), while the battery manages medium-frequency load variations, allowing the fuel cell to operate at near-constant power output within its peak efficiency zone. This hierarchical power allocation minimizes the fuel cell's response to transient demands, which would otherwise force it into low-efficiency operating regions. The increased SOC fluctuation reflects the agent's prioritization of fuel economy over energy storage cycling constraints, as the reward function structure incentivizes hydrogen consumption reduction.

As shown in Table 7, with the number of iterations increasing from 0 to 1000, fuel consumption decreased from 86.4 g to 60.0 g (a 30.5% reduction), while average system efficiency improved from 48.3% to 63.8% (a 34.6% increase). Meanwhile, the battery SOC fluctuation range expanded from 13.2% to 20.6% (a 56.1% increase). These results demonstrate that the DRL-EMS significantly reduces fuel consumption and enhances efficiency through optimized energy allocation under this operating condition, though intensified battery charge–discharge activity may accelerate battery aging. Compared to WLTP and CLTC-P conditions, the UDDS fuel economy improvement rate and SOC fluctuation increase show relatively balanced performance, indicating the algorithm's potential for universal optimization across driving cycles. However, long-term reliability assessment should still incorporate battery life models.

4. Comparative Analysis of Reinforcement Learning and Traditional Control Strategies

To systematically evaluate the operational adaptability and comprehensive performance of reinforcement learning-based energy management strategies (DRL-EMS), this section conducts a multidimensional comparative analysis of fuel consumption, system efficiency, energy distribution characteristics, and energy storage device status under three representative operating conditions (CLTC-P, UDDS, and WLTP) using simulation data. Through cross-comparison with the conventional rule-based ECMS, the technical advantages of DRL-EMS in dynamic load response and global optimization capabilities are validated.

4.1. Comparison of Hydrogen Consumption

Figures 4–6 demonstrate the comparative performance of ECMS and DRL-EMS systems. During load fluctuation periods, the load power exhibits pulse-like variations with differing peak values across algorithms. For fuel cells, ECMS maintains stable output, while DDPG and DQN dynamically adjust according to load changes—such as increasing power output during sudden surges. Power batteries discharge energy during high loads and recharge during low loads, while supercapacitors rapidly charge/discharge during load fluctuations to buffer energy. In algorithm comparison, ECMS relies on predefined rules, with fuel cells providing stable output and power batteries/supercapacitors dynamically complementing each other. DDPG demonstrates dynamic adaptability, with fuel cell power adjustments more frequent during load fluctuations and higher supercapacitor utilization in high-frequency load variations. DQN exhibits similar response speed to DDPG but with smaller adjustment amplitudes during low loads, where supercapacitor charging becomes more pronounced. In energy management strategies, ECMS emphasizes steady-state operation through rule-based energy balance maintenance, while DDPG and DQN employ reinforcement learning optimization to utilize supercapacitor characteristics for reducing high-current fluctuations in power batteries.

Hydrogen consumption research constitutes the core component of the fuel cell hybrid system technology chain. Its outcomes will directly determine the feasibility of the hydrogen economy and drive low-carbon transitions in sectors like transportation and energy storage, demonstrating significant academic value and engineering application potential. As shown in Figure 7, the fuel consumption dynamics of three algorithms—ECMS, DDPG, and DQN—are dynamically compared under three test conditions: CLTC-P, UDDS, and WLTP. The x-axis represents time (seconds), while the y-axis shows fuel consumption (grams). Under the CLTC-P condition, fuel consumption increases over time, with DDPG exhibiting the lowest consumption and slow growth, while ECMS ultimately achieves the highest consumption. In the UDDS condition (0–1400 s), initial consumption levels are

similar across algorithms, but DDPG maintains the lowest consumption later, with ECMS and DQN showing comparable but slightly higher consumption. For the WLTP condition, DDPG demonstrates lower consumption with stable growth in later stages, whereas ECMS achieves the highest consumption with a notably faster growth rate.

The hydrogen consumption per 100 km is a key evaluation parameter for fuel cell hybrid systems, as shown in Figure 8.

The bar chart reveals that the DDPG strategy demonstrates the most comprehensive performance: its total fuel consumption is the lowest among CLTC-P (75.6 g), UDDS (60.0 g), and WLTP (127.1 g), showing a 16–28% reduction compared to ECMS. It also significantly outperforms other strategies in WLTP and UDDS conditions, with 580.2 g/100 km and 560.0 g/100 km respectively. The ECMS performs the worst overall, with the highest total fuel consumption (up to 151.3 g in WLTP), yet its WLTP 100 km energy consumption is remarkably better than other strategies (722.4 g/100 km vs. DDPG's 607.1 g/100 km). This discrepancy may stem from differences in test mileage or calculation errors (e.g., in CLTC-P conditions, ECMS's total consumption of 104.8 g corresponds to 804.6 g/100 km, suggesting a test mileage of approximately 13 km, which requires verification). The DQN strategy performs moderately, but its WLTP 100 km energy consumption (682.0 g/100 km) shows significant deterioration, likely due to efficiency limitations in speed control strategies. Specific values are shown in Table 8.

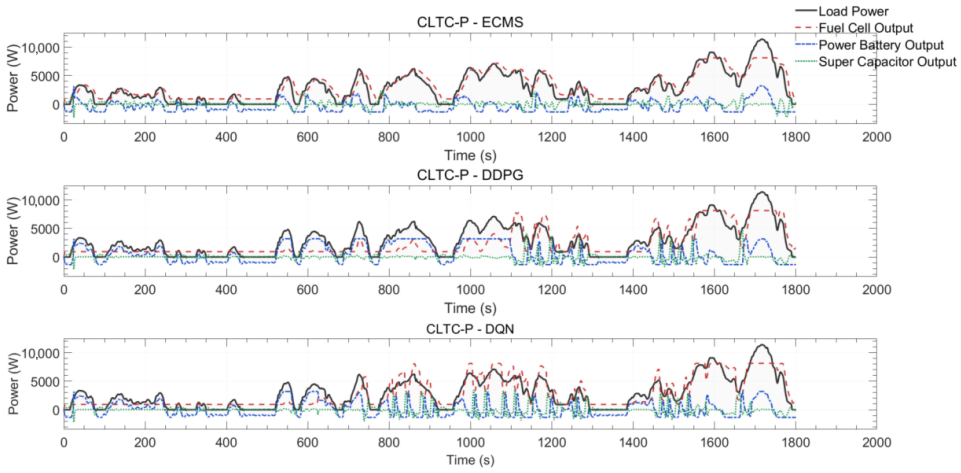

Figure 4. CLTC-P energy management performance comparison.

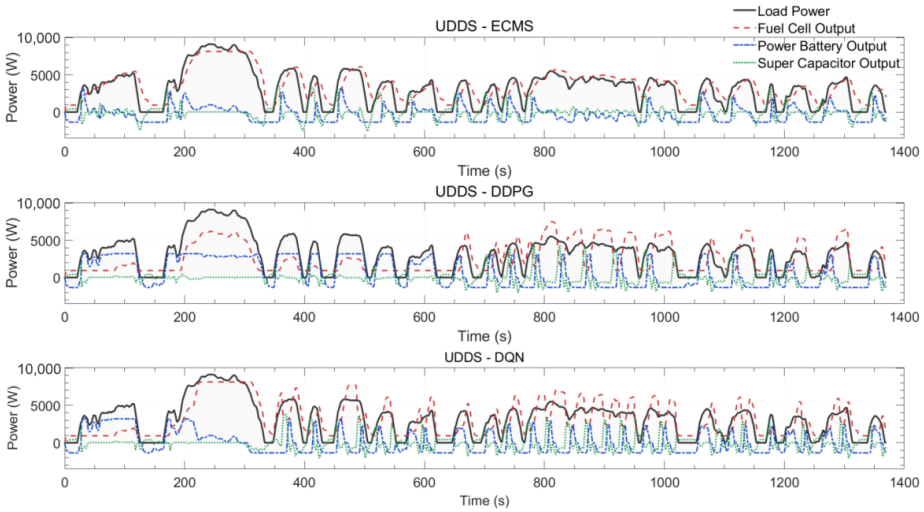

Figure 5. UDDS energy management performance comparison.

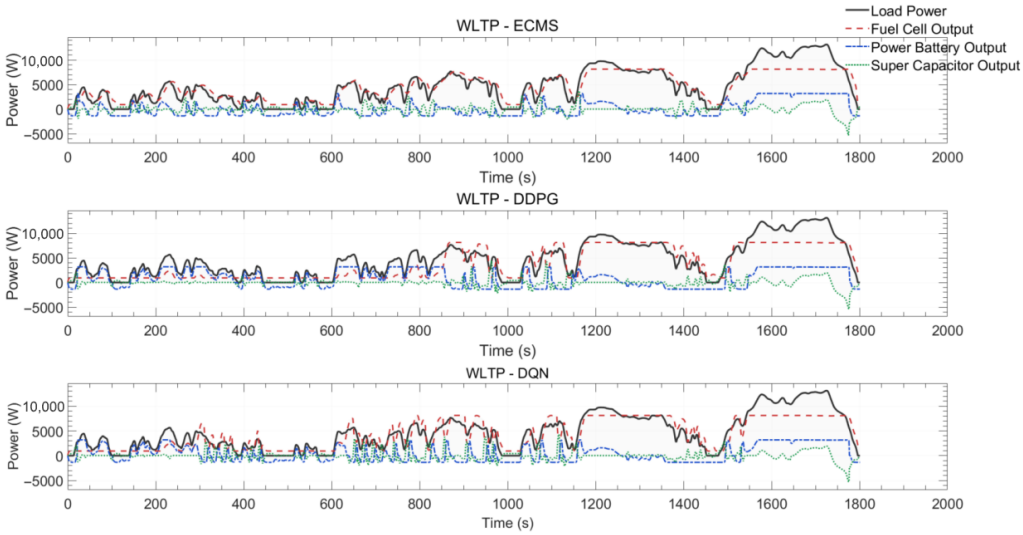

Figure 6. WLTP energy management performance comparison.

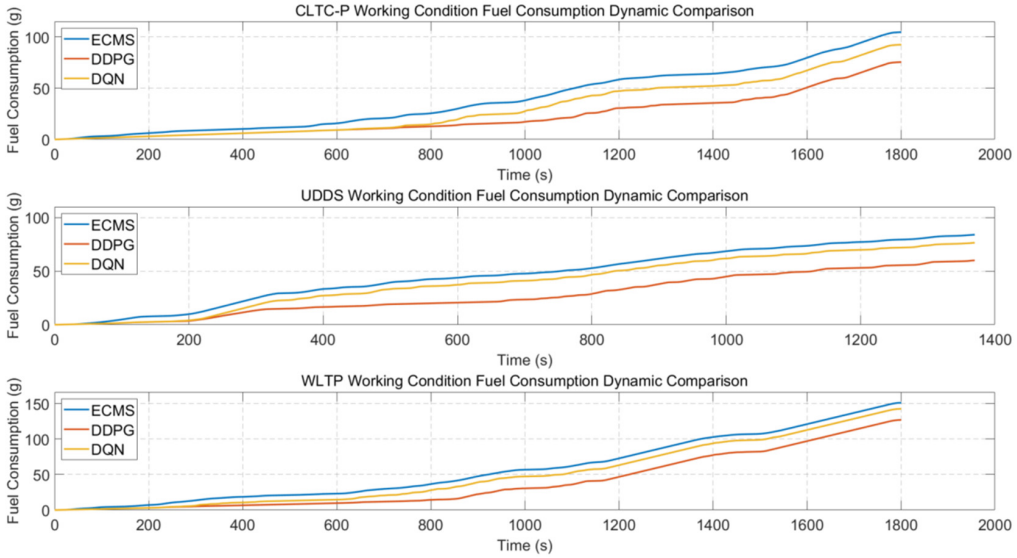

Figure 7. Dynamic fuel consumption comparison across cycles.

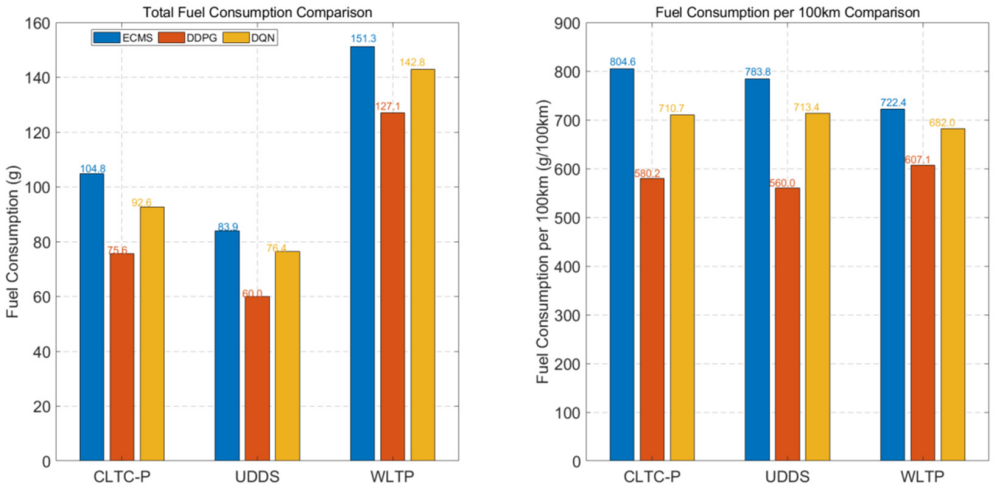

Figure 8. Total and 100 km fuel consumption comparison.

Table 8. Fuel consumption comparison across strategies.

	Fuel Consumption (g)			100 km Consumption (g/100 km)		
control strategy	CLTC-P	UDDS	WLTP	CLTC-P	UDDS	WLTP
ECMS	104.8	83.9	151.3	804.6	783.8	722.4
DDPG	75.6	60.0	127.1	580.2	560.0	607.1
DQN	92.6	76.4	142.8	710.7	713.4	682.0

4.2. System Energy Consumption and Efficiency Comparison

In fuel cell hybrid systems, while reducing hydrogen consumption remains the core objective for achieving high economic efficiency and extended range, it is equally crucial to optimize the system's overall energy consumption. The chemical energy of hydrogen fuel must be converted into electrical energy through the fuel cell stack and dynamically coupled with the charging/discharging processes of lithium batteries and supercapacitors. The comprehensive energy efficiency of the system is determined by multiple factors: conversion efficiency across various stages (e.g., fuel cell polarization losses, DC/DC converter losses, thermal losses from energy storage media), dynamic power distribution coordination through energy management strategies (e.g., avoiding high-frequency shallow charging of lithium batteries or overcharging of supercapacitors to prevent efficiency loss), and parasitic power consumption of auxiliary systems (e.g., air compressors, cooling devices, BMS/EMS). Only through multi-dimensional collaborative optimization at the system level can truly efficient, economical, and sustainable operation be achieved, as illustrated in Figure 9.

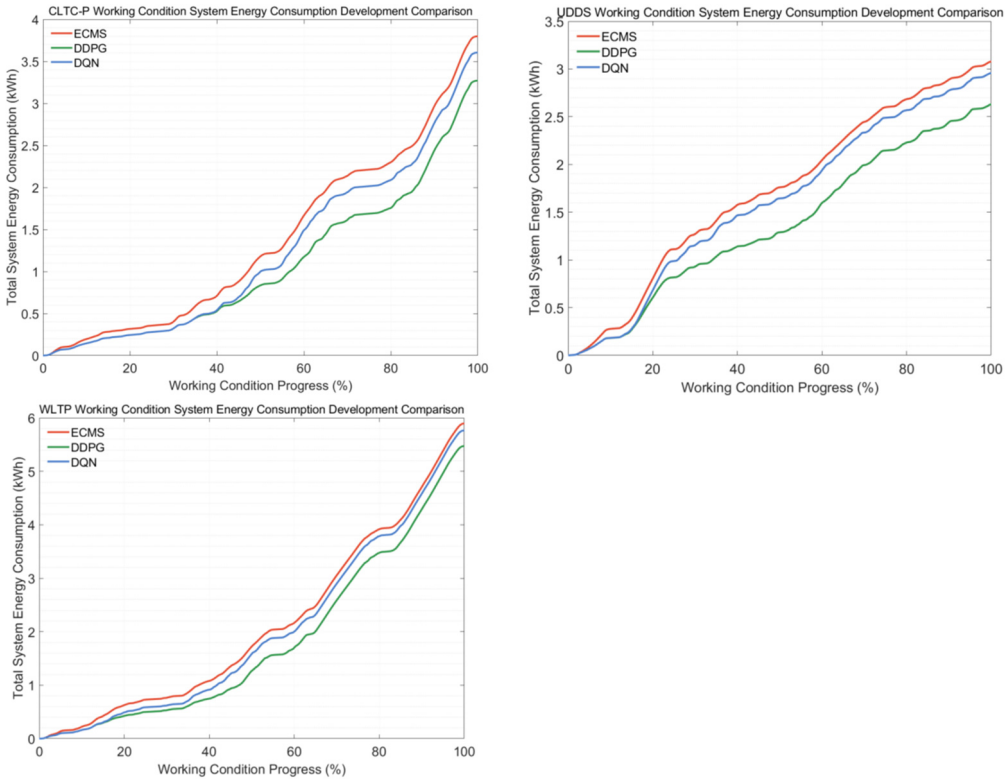

Figure 9. Dynamic system energy consumption comparison.

Under the CLTC-P, UDDS, and WLTP test conditions, the total system energy consumption of ECMS, DDPG, and DQN methods all showed an upward trend with the progression of the test conditions, though their specific energy consumption performance varied. In the CLTC-P condition, the ECMS curve consistently remained above the others, demonstrating

the highest total energy consumption, while the DDPG curve stayed below with relatively lower energy consumption. For the UDDS condition, the energy consumption trends of ECMS and DQN were similar in the early stages, but ECMS slightly outperformed DQN later, whereas DDPG exhibited the smallest growth in energy consumption, showing a clear overall advantage. Under the WLTP condition, the energy consumption curves of ECMS and DQN were closely similar in the early stages, but ECMS showed slightly higher energy consumption later, while DDPG maintained a low growth trend. Overall, the DDPG method demonstrated superior energy consumption control across all three conditions, with relatively lower total system energy consumption, whereas ECMS exhibited higher energy consumption in most conditions. The comprehensive comparison results are shown in Figure 10.

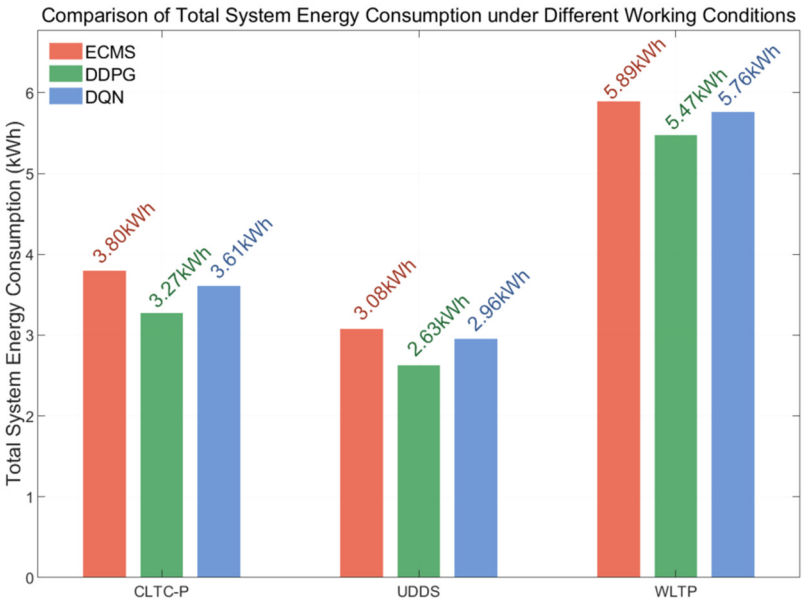

Figure 10. Final system energy consumption comparison.

The overall results demonstrate that DDPG exhibits significantly lower total energy consumption across all operating conditions compared to ECMS and DQN, demonstrating superior energy-saving performance. The specific numerical values are presented in Table 9.

Table 9. Total system energy consumption comparison.

Operating Mode	ECMS (kWh)	DDPG (kWh)	DQN (kWh)
CLTC-P	3.80	3.27	3.61
UDDS	3.08	2.63	2.94
WLTP	5.89	5.47	5.76

In power systems, energy efficiency serves as a critical nexus between economic costs, environmental benefits, and system performance. By reducing energy consumption per task unit, it directly lowers operational expenses (e.g., equipment electricity bills and vehicle charging costs). Lower energy consumption—particularly through reduced fossil fuel usage—effectively cuts carbon emissions and pollutant releases, supporting green and low-carbon goals. From a performance perspective, high energy efficiency extends equipment range, enhances continuous operation capability, and improves stability and reliability. Additionally, it plays a vital role in alleviating energy supply–demand imbalances and promoting renewable energy applications like solar power in power systems, driving the transition toward cleaner and more sustainable energy structures. This achieves coordinated development across economic, environmental, and energy systems. As shown

in Figure 11, the DDPG strategy demonstrates the lowest energy efficiency values across three test conditions (CLTC-P, UDDS, and WLTP), indicating superior energy efficiency. In contrast, the ECMS generally exhibits higher energy efficiency across all conditions but relatively poorer energy consumption performance.

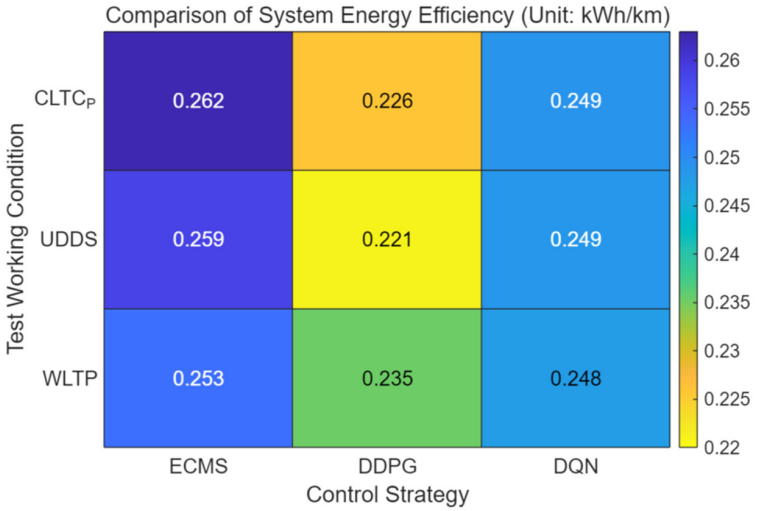

Figure 11. System energy efficiency comparison.

Figure 12 presents a comparative analysis of system efficiency across different operating conditions and systems under CLTC-P, UDDS, and WLTP standards. The ECMS, DDPG, and DQN strategies demonstrate distinct temporal trends in efficiency performance. Under CLTC-P conditions, the DQN strategy initially achieves high efficiency (around 70%), followed by the DDPG strategy stabilizing at a higher level (approximately 60–70%), while the ECMS remains consistently low (50–60%). In the UDDS scenario, the DQN strategy initially outperforms (exceeding 80%), but subsequently declines, whereas the DDPG strategy maintains stability at a high efficiency range (60–70%) and the ECMS stays at a relatively low level (50–60%). For the WLTP conditions, the DDPG strategy maintains high efficiency (60–70%) after initial fluctuations, while the DQN strategy shows some decline and the ECMS remains the lowest. Overall, the DDPG strategy demonstrates superior efficiency performance across most stages and exhibits relatively outstanding stability in various operating conditions, whereas the ECMS generally performs weaker.

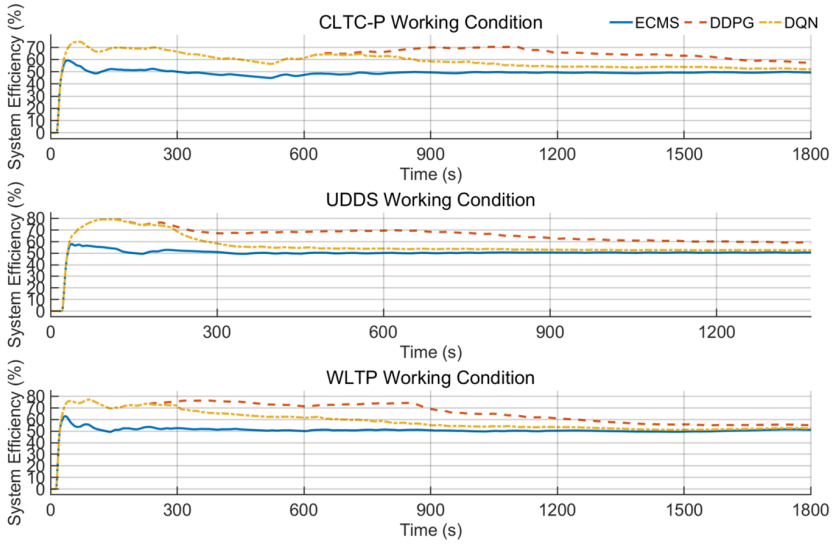

Figure 12. System efficiency dynamics across cycles.

4.3. Stability Analysis of State of Charge (SOC) in ESSs

Figure 13 demonstrates the dynamic control performance of the ECMS, DQN, and DDPG strategies on lithium battery State of Charge (SOC) under three operating conditions: CLTC-P, UDDS, and WLTP. Overall, the ECMS control strategy demonstrates superior performance in maintaining battery SOC, exhibiting higher mean values, lower standard deviations, and relatively higher minimum values, thereby better protecting the battery and sustaining higher energy levels. The DDPG control strategy, however, shows significant SOC fluctuations, suggesting the need for further optimization to mitigate its impact on battery lifespan. Detailed statistical comparisons are presented in Table 10.

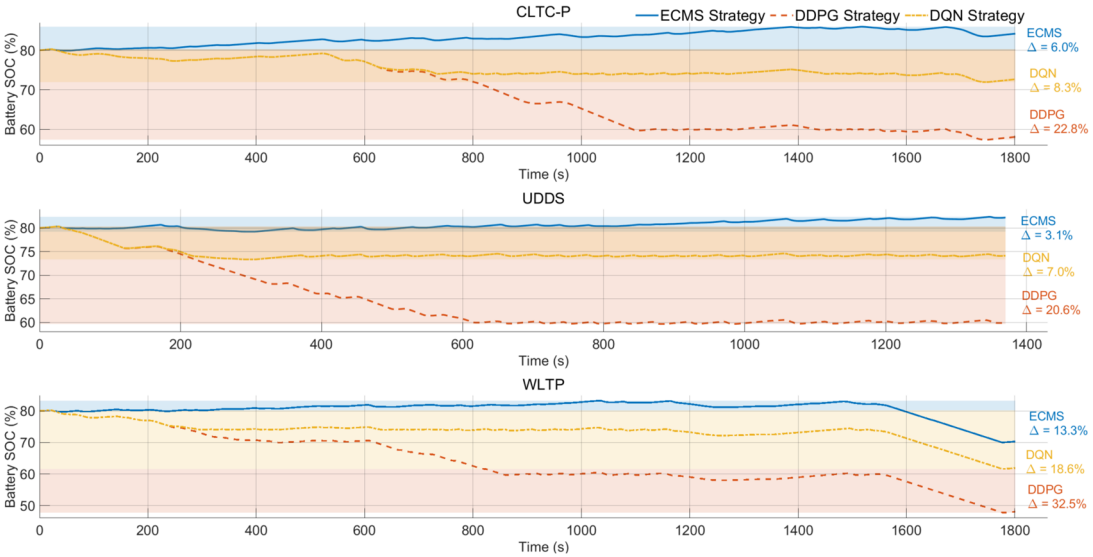

Figure 13. Lithium battery SOC fluctuations under different strategies.

Table 10. Lithium battery SOC statistics.

Operating Mode	Control Strategy	Average	Standard	Minimum	Maximum
UDDS	DDPG	64.50	6.47	59.69	80.32
UDDS	DQN	74.60	1.40	73.36	80.32
UDDS	ECMS	80.65	0.84	79.22	82.36
CLTC-P	DDPG	68.63	8.31	57.44	80.26
CLTC-P	DQN	75.57	2.11	71.97	80.26
CLTC-P	ECMS	83.21	1.76	79.90	85.94
WLTP	DDPG	64.30	8.07	47.76	80.24
WLTP	DQN	73.56	3.31	61.64	80.24
WLTP	ECMS	80.69	2.63	70.05	83.30

Under various operating conditions, the average State of Charge (SOC) of lithium batteries under the ECMS control strategy consistently remains high, demonstrating its superior performance in maintaining battery capacity. In contrast, the average SOC under DDPG control strategy is relatively lower, suggesting faster battery consumption under this approach. For instance, under the WLTP test cycle, the ECMS achieves an average SOC of 80.68%, while the DDPG strategy only reaches 64.30%.

Standard deviation measures data dispersion. The ECMS control strategy typically exhibits a smaller standard deviation, indicating stable SOC fluctuations and consistent battery charge–discharge behavior. In contrast, the DDPG control strategy demonstrates a larger standard deviation (e.g., 8.31 under the CLTC-P test conditions), suggesting greater SOC fluctuations in lithium batteries under this strategy.

Figure 14 demonstrates the dynamic control performance of ECMS, DQN, and DDPG strategies on supercapacitor State of Charge (SOC) under three operating conditions: CLTC-P, UDDS, and WLTP. The specific numerical values are presented in Table 11.

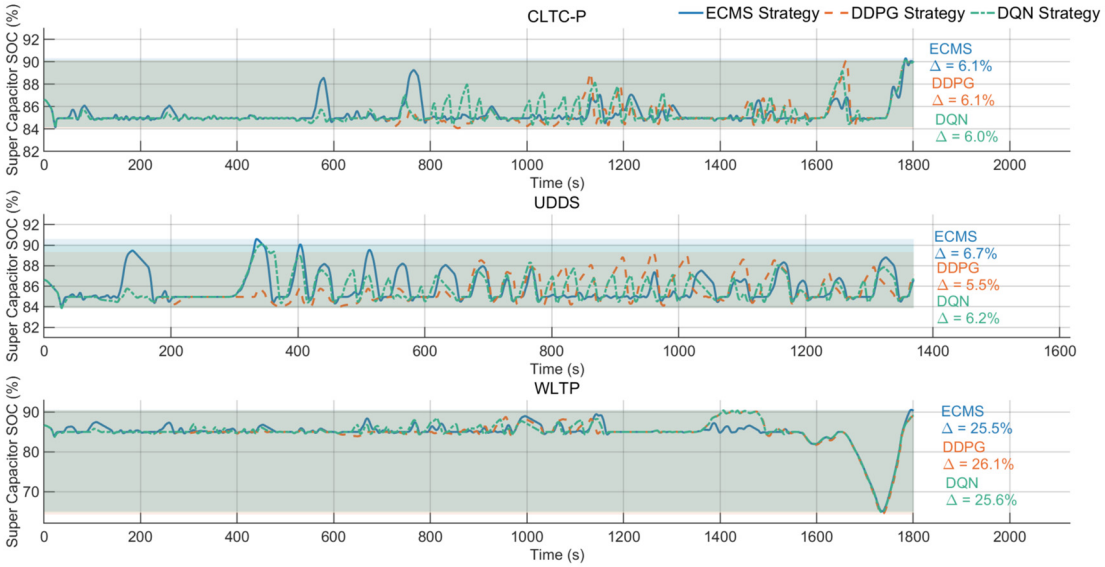

Figure 14. Supercapacitor SOC fluctuations under different strategies.

Table 11. Supercapacitor SOC statistics.

Operating Mode	Control Strategy	Average	Standard	Minimum	Maximum
UDDS	DDPG	85.59	1.22	83.87	89.33
UDDS	DQN	85.73	1.17	83.87	90.07
UDDS	ECMS	85.87	1.38	83.88	90.59
CLTC-P	DDPG	85.25	0.98	84.07	90.13
CLTC-P	DQN	85.37	0.95	84.17	90.13
CLTC-P	ECMS	85.36	0.91	84.17	90.31
WLTP	DDPG	84.66	3.56	64.24	90.37
WLTP	DQN	84.84	3.48	64.77	90.37
WLTP	ECMS	84.69	3.29	65.02	90.52

Under various operating conditions, the average State of Charge (SOC) values of supercapacitors under different control strategies show significant similarity. Overall, the ECMS control strategy demonstrates slightly higher average values in both UDDS and CLTC-P conditions compared to other strategies. In WLTP conditions, its average value closely matches the DQN strategy while slightly exceeding the DDPG strategy. These findings suggest that the ECMS may exhibit certain advantages in maintaining supercapacitor charge levels, though the benefits are not particularly pronounced.

Standard deviation measures the dispersion of data. Under the WLTP test conditions, all strategies exhibit relatively high standard deviations, indicating significant SOC fluctuations in supercapacitors. The DDPG strategy demonstrates the highest standard deviation of 3.56 under WLTP conditions, reflecting substantial variations in charge–discharge behavior. In contrast, both UDDS and CLTC-P conditions show lower standard deviations, demonstrating more stable supercapacitor SOC performance.

5. Summary

This chapter focuses on the application and optimization of the Deep Deterministic Policy Gradients (DDPG) algorithm in energy management for fuel cell hybrid vehicles. Research demonstrates that DDPG achieves dynamic global optimization of power allo-

cation between fuel cells and energy storage systems through its Actor–Critic network architecture and continuous action space design. Under three typical operating conditions (WLTP, CLTC-P, and UDDS), the DDPG strategy achieves hydrogen consumption of 607.1 g/100 km, 580.2 g/100 km, and 560.0 g/100 km respectively, representing a maximum 28% reduction compared to traditional ECMS strategies. The system average efficiency improves by approximately 38.9%, reaching 66% under WLTP conditions. The core breakthrough lies in utilizing supercapacitors' high-frequency response characteristics to mitigate load fluctuations, combined with dynamic adjustment of fuel cells' high-efficiency operating range (a 15% increase in load ratio from 40% to 80%), significantly reducing inefficient start–stop cycles and power fluctuations. However, the SOC fluctuation range of lithium batteries under the DDPG strategy expands by 32.5% compared to ECMS (WLTP conditions), necessitating optimization of the reward function's SOC deviation penalty term to balance energy efficiency and battery life degradation.

It should be noted that the current study primarily focuses on short-to-medium-term energy management optimization, while the long-term durability degradation of the fuel cell stack under dynamic driving conditions has not been explicitly incorporated into the optimization framework. In practice, frequent load transients, start–stop events, and suboptimal operating points can accelerate membrane degradation and catalyst layer deterioration in proton exchange membrane fuel cells (PEMFCs). Although the proposed DDPG strategy indirectly mitigates degradation by maintaining the fuel cell within the high-efficiency load region (40–80%) and reducing unnecessary start–stop events through supercapacitor power buffering, a quantitative degradation model has not been integrated into the current reward function.

For long-term durability considerations under dynamic driving conditions, health state estimation frameworks—such as those proposed for vehicular PEM fuel cell stacks operating under real-world transient conditions—may offer predictive insights regarding performance degradation trends. In future work, such degradation indicators (e.g., voltage decay rate, membrane resistance increase) could be incorporated as additional penalty terms in the DDPG reward function, enabling the agent to account for cumulative degradation effects during the training process. This would allow for a more holistic multi-objective optimization that simultaneously balances instantaneous energy efficiency, SOC stability, and fuel cell long-term lifespan.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L.; Software, Y.L.; Validation, X.H.; Formal analysis, X.H.; Investigation, X.H.; Resources, Y.L.; Data curation, M.L. and X.H.; Writing—original draft, M.L.; Writing—review & editing, M.L.; Visualization, Y.L.; Supervision, X.H.; Project administra-tion, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by Naval University of Engineering.

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 conflict of interest.

References

1. Li, X.; Ye, T.; Meng, X.; He, D.; Li, L.; Song, K.; Jiang, J.; Sun, C. Advances in the Application of Sulfonated Poly(Ether Ether Ketone) (SPEEK) and Its Organic Composite Membranes for Proton Exchange Membrane Fuel Cells (PEMFCs). *Polymers* **2024**, *16*, 2840. [CrossRef]
2. Wang, Y.; Wang, L.; Li, M.; Chen, Z. A review of key issues for control and management in battery and ultra-capacitor hybrid energy storage systems. *eTransportation* **2020**, *4*, 100064. [CrossRef]

3. Meng, X.; Sun, C.; Mei, J.; Tang, X.; Hasanien, H.M.; Jiang, J.; Fan, F.; Song, K. Fuel cell life prediction considering the recovery phenomenon of reversible voltage loss. *J. Power Sources* **2025**, *625*, 235634. [CrossRef]
4. Amphlett, J.C.; Baumert, R.M.; Mann, R.F.; Peppley, B.A.; Roberge, P.R.; Harris, T.J. Performance modeling of the ballard mark iv solid poljoner electrolyte fuel cell: I. mechanistic mode development. *J. Electrochem. Soc.* **1995**, *142*, 1. [CrossRef]
5. Truc, N.T.; Ito, S.; Fushinobu, K. Numerical and experimental investigation on the reactant gas crossover in a pem fuel cell. *Int. J. Heat Mass Transf.* **2018**, *127*, 447–456. [CrossRef]
6. Wu, X.; Zou, P.; Fu, J.; Deng, P.; Yang, Y.; Cai, Y.; Zeng, Z. Research progress on energy management strategies for fuel cell electric vehicle power systems. *J. Tsinghua Univ. (Nat. Sci. Ed.)* **2020**, *39*, 89–96.
7. Saygili, Y.; Eroglu, I.; Kincal, S. Model based temperature controller development for water cooled pem fuel cell systems. *Int. J. Hydrogen Energy* **2015**, *40*, 615–622. [CrossRef]
8. Wang, Y.; Tian, J.; Sun, Z.; Wang, L.; Xu, R.; Li, M.; Chen, Z. A comprehensive review on battery modeling and state estimation approaches for advanced battery management systems. *Renew. Sustain. Energy Rev.* **2020**, *131*, 110015. [CrossRef]
9. Grandjean, T.R.; Li, L.; Odio, M.X.; Widanage, W.D. Global sensitivity analysis of the single particle lithium-ion battery model with electrolyte. In *Proceedings of the 2019 IEEE Vehicle Power and Propulsion Conference (VPPC), Hanoi, Vietnam, 14–17 October 2019*; IEEE: New York, NY, USA, 2019; pp. 1–7.
10. Doyle, M.; Fuller, T.F.; Newman, J. Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell. *J. Electrochem. Soc.* **1993**, *140*, 1526. [CrossRef]
11. Han, X.; Ouyang, M.; Lu, L.; Li, J. Simplification of physics-based electrochemical model for lithium ion battery on electric vehicle part II: Pseudo-two-dimensional model simplification and state of charge estimation. *J. Power Sources* **2015**, *278*, 814–825. [CrossRef]
12. Ding, Q.; Wang, Y.; Chen, Z. Parameter identification of reduced-order electrochemical model simplified by spectral methods and state estimation based on square root cubature kalman filter. *J. Energy Storage* **2022**, *46*, 103828. [CrossRef]
13. Liaw, B.Y.; Nagasubramanian, G.; Jungst, R.G.; Doughty, D.H. Modeling of lithium ion cells—A simple equivalent circuit model approach. *Solid State Ion.* **2004**, *175*, 835–839.
14. Freeborn, T.J.; Maundy, B.; Elwakil, A.S. Fractional-order models of supercapacitors, batteries and fuel cells: A survey. *Mater. Renew. Sustain. Energy* **2015**, *4*, 9. [CrossRef]
15. Yang, Q.; Xu, J.; Cao, B.; Li, X. A simplified fractional order impedance model and parameter identification method for lithium-ion batteries. *PLoS ONE* **2017**, *12*, e0172424. [CrossRef]
16. Lü, X.; Qu, Y.; Wang, Y.; Qin, C.; Liu, G. A comprehensive review on hybrid power system for PEMFC-HEV: Issues and strategies. *Energy Convers. Manag.* **2018**, *12*, 73–91. [CrossRef]
17. Hua, Z.; Zheng, Z.; Pahon, E.; Péra, M.-C.; Gao, F. A review on lifetime prediction of proton exchange membrane fuel cells system. *J. Power Sources* **2022**, *529*, 231256. [CrossRef]
18. Liu, B.; Wei, X.; Sun, C.; Wang, B.; Huo, W. A controllable neural network-based method for optimal energy management of fuel cell hybrid electric vehicles. *Int. J. Hydrogen Energy* **2024**, *55*, 1371–1382. [CrossRef]
19. Feng, J.; Han, Z.; Wu, Z.; Li, M. Approximate Optimal Energy Management with a High-Precision Vehicle Speed Prediction Algorithm. *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.* **2022**, *238*, 774–787. [CrossRef]
20. Ma, R.; Yang, T.; Breaz, E.; Li, Z.; Briois, P.; Gao, F. Data driven proton exchange membrane fuel cell degradation predication through deep learning method. *Appl. Energy* **2018**, *231*, 102–115. [CrossRef]
21. Li, Q.; Meng, X.; Gao, F.; Zhang, G.; Chen, W.; Rajashekara, K. Reinforcement Learning Energy Management for Fuel Cell Hybrid System: A Review. *IEEE Ind. Electron. Mag.* **2022**, *17*, 45–54. [CrossRef]
22. Ganesh, A.H.; Xu, B. A Review of Reinforcement Learning Based Energy Management Systems for Electrified Powertrains: Progress, Challenge, and Potential Solution. *Renew. Sustain. Energy Rev.* **2022**, *154*, 111833. [CrossRef]
23. 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]
24. Deng, Z.; Chan, S.H.; Chen, Q.; Liu, H.; Zhang, L.; Zhou, K.; Tong, S.; Fu, Z. Efficient degradation prediction of PEMFCs using ELM-AE based on fuzzy extension broad learning system. *Appl. Energy* **2023**, *331*, 22619. [CrossRef]
25. Mei, J.; Meng, X.; Tang, X.; Li, H.; Hasanien, H.; Alharbi, M.; Dong, Z.; Shen, J.; Sun, C.; Fan, F.; et al. An Accurate Parameter Estimation Method of the Voltage Model for Proton Exchange Membrane Fuel Cells. *Energies* **2024**, *17*, 2917. [CrossRef]
26. Liu, J.W.; Li, Q.; Chen, W.R. Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks. *Int. J. Hydrogen Energy* **2019**, *44*, 115470. [CrossRef]
27. Tao, Z.; Zhang, C.; Xiong, J.L. Evolutionary gate recurrent unit coupling convolutional neural network and improved manta ray foraging optimization algorithm for performance degradation prediction of PEMFC. *Appl. Energy* **2023**, *336*, 120821. [CrossRef]
28. He, K.; Mao, L.; Yu, J.; Huang, W.; He, Q.; Jackson, L. Long-Term Performance Prediction of PEMFC Based on LASSO-ESN. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3511611. [CrossRef]

29. Ahmed, T.; Zhang, H.; Yan, B. A Review on Renewable Energy and Electricity Requirement Forecasting Models for Smart Grid and Buildings. *Sustain. Cities Soc.* **2020**, *55*, 102052. [CrossRef]
30. Chen, Z.; Hu, H.; Wu, Y.; Zhang, Y.; Li, G.; Liu, Y. Stochastic Model Predictive Control for Energy Management of Power-Split Plug-in Hybrid Electric Vehicles Based on Reinforcement Learning. *Energy* **2020**, *211*, 118931. [CrossRef]
31. Huo, W.; Chen, D.; Tian, S.; Li, J.; Zhao, T.; Liu, B. Lifespan-Consciousness and Minimum-Consumption Coupled Energy Management Strategy for Fuel Cell Hybrid Vehicles via Deep Reinforcement Learning. *Int. J. Hydrogen Energy* **2022**, *47*, 24026–24041. [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.