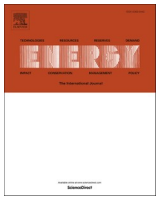
Energy 338 (2025) 138782



Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy



Safe reinforcement learning-based energy management for fuel cell hybrid electric aircraft with longevity considerations

Yajing Xiao [a], Jinning Zhang [a,*], Harold S. Ruiz [a], Ioannis Roumeliotis [b], Xin Zhang [c]

[a] Green Energy and Transportation Research Group, School of Engineering, University of Leicester, Leicester, LE1 7RH, UK
[b] Centre for Propulsion and Thermal Power Engineering, Cranfield University, Bedford, MK43 0AL, UK
[c] Electrical Machines and Drives Research Group, School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, S10 2TN, UK

ARTICLE INFO

Handling Editor: X Ou

Keywords:
Electric aircraft
Energy management strategy
Safe reinforcement learning
Multi-objective optimization

ABSTRACT

Fuel Cell Hybrid Electric Aircraft (FCHEA) represent a promising solution for decarbonizing short-to medium-range aviation. However, the hybrid-electric architecture introduces increased control complexity and poses challenges in ensuring component longevity and operational safety. Although reinforcement learning (RL)-based energy management strategies (EMS) have been explored in ground vehicle application, they often prioritize fuel efficiency while neglecting component degradation and safety-critical constraints, both of which are vital for the reliability of electric aviation. This study presents a Longevity-Conscious Safe Energy Management Strategy (LC-SEMS) to minimize operational and degradation-related costs over long-term use, while ensuring the satisfaction of multi-type constraint. The strategy is implemented within a multidisciplinary simulation framework that integrates propulsion, aerodynamics, hybrid powertrain, and flight dynamics models for mission-level evaluation. The EMS problem is formulated as a Constrained Markov Decision Process (CMDP) incorporating physical, cumulative, and instantaneous constraints. Instantaneous safety is enforced via an adaptive shielding mechanism that leverages a pretrained transition model to detect potential constraint violations and applies minimal corrective actions without interfering with policy learning. The proposed strategy is validated on a simulated FCHEA retrofitted from the NASA X-57 Maxwell, achieving fast convergence and strict constraint adherence across multi-mission scenarios. It achieves a 26.96 % reduction in depreciation cost compared to baseline RL-based EMS, with a minimal 4.21 % performance gap relative to the globally optimal Dynamic Programming (DP) benchmark, demonstrating its adaptability and robustness under uncertain and unseen mission scenarios.

1. Introduction

The aviation industry contributes significantly to global carbon emissions, necessitating urgent efforts to achieve decarbonization by 2050 [1]. Among emerging technologies, electrified propulsion offers a promising path toward zero-emission and low-noise flight [2]. However, purely battery-electric configurations are limited by short range and prolonged charging time [3]. On the other hand, fuel cell (FC)-only electric systems face cold-start delays and slow transient response, making them unsuitable for highly dynamic power demand characteristic of take-off phase [4]. Moreover, operating FCs under fluctuating loads not only undermines system safety but also accelerates component degradation [5]. To address these limitations, hybridizing fuel cells with batteries has emerged as a technically feasible approach to overcoming these challenges, as it leverages the high energy density of FCs and the fast power response of batteries. While such hybrid architectures can significantly enhance system performance and reliability, they introduce new complexities. Specifically, the energy management system (EMS) must efficiently coordinate power distribution, maintain battery state-of-charge (SoC), and mitigate component aging, while simultaneously ensuring safe operation under uncertain mission conditions.

EMSs can be broadly categorized into rule-based, optimization-based, and reinforcement learning (RL)-based approaches [6]. Rule-based EMSs, such as State machine [7] and Fuzzy Logic (FL) [8], offer simplicity and real-time applicability, but their reliance on expert-defined thresholds limits adaptability under dynamic and uncertain flight conditions [9]. Optimization-based EMSs such as Equivalent Consumption Minimization Strategy (ECMS) [10] and Model Predictive Control (MPC) [11] have been introduced to improve fuel efficiency by leveraging system models and future predictions have been

* Corresponding author.
E-mail address: jz388@leicester.ac.uk (J. Zhang).

https://doi.org/10.1016/j.energy.2025.138782
Received 4 August 2025; Received in revised form 13 September 2025; Accepted 3 October 2025
Available online 3 October 2025
0360-5442/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

proposed to achieve better performance in multi-objective scenarios. However, their effectiveness hinges on the availability of accurate system models and precise forecasts of future mission trajectories, which are often difficult to obtain in high-dimensional, uncertain, and dynamic flight environments [12]. As a result, global methods such as Dynamic Programming (DP) serve as a valuable reference for evaluating the upper-bound performance of optimal EMS algorithms, rather than functioning as viable real-time solutions [13]. To address these limitations, RL-based EMSs provide a model-free alternative that is particularly well-suited for complex and uncertain flight scenarios [14]. Early applications were dominated by value-based algorithms, which estimate action-value functions in discrete control spaces. For example, a dual-reward Q-learning EMS was shown to improve average system efficiency to 52 % in hybrid electric vehicles [15], while degradation-aware Q-learning achieved cost performance comparable to offline optimization, with only a 4.78 % deviation [16]. However, the inherent limitations of value-based methods in handling continuous action spaces have prompted a shift toward policy-based approaches. Zhang et al. [17] enhanced the exploration-exploitation balance of Deep Deterministic Policy Gradient (DDPG) by integrating a Locality Sensitive Hashing mechanism, resulting in improved learning performance in continuous control domains. Mazzi et al. [18] proposed an improved soft actor-critic framework that demonstrated increased robustness and fuel savings in hybrid energy systems. These advancements highlight the strong potential of RL-based EMSs to deliver adaptive, efficient, and scalable solutions for complex hybrid energy management scenarios.

However, conventional RL training is based on trial-and-error and reward feedback [19], whose inherently exploratory nature often leads to safety constraint violations, both during training and deployment. From a control-theoretic perspective, safety constraints can be categorized into instantaneous (hard) constraints, which must be satisfied at each time step, and cumulative (soft) constraints, which tolerate occasional violations as long as the overall accumulation remains acceptable [20]. Table 1 summarizes the mainstream safe RL-based EMSs and methodologies.

In most RL-based EMS, constraints are implicitly handled through reward shaping, which modifies the reward function to penalize unsafe behaviour [21]. A representative subclass is Lagrangian relaxation, which introduces Lagrange multipliers to reformulate the constrained problem into an unconstrained dual. For example, Zhang et al. [22] employed this technique to integrate physical constraints into the EMS for hybrid electric vehicles, achieving over 95 % of the offline DP benchmark in energy efficiency. Similarly, Huang et al. [23] incorporated SoC penalties directly into the reward function to discourage constraint violations. While conceptually simple and easy to implement, these methods suffer from limited temporal resolution due to the cumulative nature of reward signals. Consequently, they lack guarantees of state-wise safety and often require extensive tuning to balance reward and penalty components. In contrast, projection-based Safe RL methods explicitly enforce constraints through the projection actions onto a feasible set. Zhang et al. [24] introduced such an approach for autonomous vehicle motion planning, ensuring that the state trajectory remains within a predefined safe set. Similarly, Zhang et al. [25] developed a projection-embedded control scheme for distributed energy systems, achieving decentralized optimality while maintaining strict safety compliance. However, these methods incur high computational costs, as projection must be performed at every decision step and state. Since all actions are treated as potentially unsafe, the algorithm applies projection universally, leading to redundant computation even when no violation occurs. Moreover, non-differentiable projection operations, such as hard clipping or quadratic programming, disrupt gradient flow and can cause vanishing gradients during early training stages [26], ultimately hindering learning speed and stability. A third class of methods, Lyapunov-based approaches, which aim to guarantee system stability via theoretical formulations. For example, Hao et al. [27] applied this approach to microgrid control by refining policies to eliminate unsafe trajectories. While theoretically grounded, these methods are challenging to scale to high-dimensional systems, as constructing valid Lyapunov functions often requires accurate analytical models, which are impractical in complex environments. Although finer-grained Safe RL categories have been proposed, such as risk-sensitive methods, shielding techniques, or control barrier functions, many of these can be viewed as special cases of the three categories summarized here, and are less applicable to the aircraft EMS setting [28].

While Safe RL has received increasing attention in domains such as smart grids [29] and ground vehicle energy systems [30], these settings differ fundamentally from aviation. In smart grids, power demand and generation are relatively periodic and predictable, while ground vehicles operate under less aggressive dynamics with moderate and gradual power transients primarily governed by friction. In contrast, aircraft must deliver substantial propulsion within seconds during take-off, rapidly transitioning from idle to peak power. This demand is further intensified by sharp load fluctuations during manoeuvres and unpredictable external disturbances, all within a tightly coupled, high-dimensional control system. Moreover, safety constraints in aircraft energy management are structurally diverse, encompassing physical constraints derived from system models, degradation-related constraints that evolve over time, and instantaneous constraints requiring each control action to remain within safe bounds under uncertainty. This heterogeneity renders constraint handling significantly more complex than in conventional applications and poses fundamental challenges for existing safe RL methods, which typically assume uniform or softly enforced constraints. These factors collectively underline the necessity for a specialised safety-aware EMS framework in the aviation. Beyond safety, degradation-aware control has gained growing importance in EMS-related multi-objective optimization studies. Yadav et al. [31] and Abdelsattar et al. [32] explored degradation modelling in microgrid EMS involving hybrid energy storage systems. These studies typically adopt simplified approaches by fixed degradation rates or average lifetime cost allocation over the system's lifespan. However, such static formulations fail to capture the dynamic nature of degradation under varying operational conditions, thereby limiting the effectiveness of the optimization and its coupling with control strategy. In the vehicle domain, Zhou et al. [33] applied aging-aware modifications to

Table 1
Summary of safe RL-based energy management strategy methodologies.

Method Class	Constraint Type	Enforcement Principle	Representative Techniques	Advantages	Limitations
Reward Shaping	Soft (cumulative)	Reformulate an unconstrained problem, Penalizes unsafe actions via reward modification.	Lagrangian Relaxation, Primal-Dual Methods	Simple implementation; compatible with standard RL	No guarantee of stepwise safety; requires extensive tuning
Projection-Based	Hard (instantaneous)	Project actions onto feasible set at each step	Constrained Policy Optimization (CPO), Projection-Based CPO, Conservative Policy Updates	Ensures constraint satisfaction at every time step	High computational cost; vanishing gradients; non-differentiable projection
Lyapunov-Based	Hard (stability-driven)	Ensures stability using Lyapunov conditions	Lyapunov function	Theoretically grounded; ensures long-term stability	Requires accurate system model; difficult to scale to high-dimensional systems

the Equivalent Consumption Minimization Strategy (ECMS) in hybrid electric vehicles. However, ECMS still fundamentally focuses on energy efficiency, treating aging effects as auxiliary costs rather than as co-primary objectives in the optimization. This leads to a lack of explicit trade-off mechanisms between energy efficiency and component lifespan. Moreover, many existing works, such as [34,35], focus on the degradation of a single component, most commonly the battery, without accounting for multi-source interactions in hybrid architectures. Such interactions may not only alter the load distribution but also affect each other's aging rates, leading to cross-component degradation effects that are currently underexplored. This limitation becomes particularly critical in the context of all-electric or hybrid-electric aviation propulsion systems. The service lives of key power components are relatively limited, with lithium-ion batteries lasting about 3000 cycles [36] and fuel cells about 5000 h [37], in contrast to traditional turbine engines that typically exceed 100,000 flight hours [38]. This mismatch leads to high lifecycle replacement costs, making longevity-aware EMS design essential for economic feasibility. There is still a lack of unified EMS frameworks that integrate real-time control with degradation-aware multi-objective optimization. In particular, few studies formulate lifespan explicitly as a primary optimization criterion alongside energy and cost, especially under dynamic operating conditions. This highlights a critical research gap and motivates our ongoing efforts to develop coordinated control strategies that jointly optimize energy performance, component health, and operational safety in a unified framework.

To bridge the afore mentioned research gaps, this paper proposes a longevity-conscious safe energy management strategy (LC-SEMS) for FCHEA aimed at minimizing operational costs and longevity-related degradation cost while ensuring satisfaction of multi-type safety constraints. The main contributions of this article are as follows.

1) A multidisciplinary EMS environment is constructed by propulsion, aerodynamics, hybrid powertrain, and flight dynamics models. Scenario-dependent degradation modelling, informed by flight operational conditions, is incorporated for both the fuel cell and battery to support joint optimization of fuel efficiency and component longevity. Moreover, wind disturbances are also included to support robust training under uncertainties.
2) A Safe RL-based EMS strategy is formulated within a Constrained Markov Decision Process (CMDP) framework to jointly optimize hydrogen consumption, component degradation, and SoC regulation. Physical, cumulative, and instantaneous constraints are systematically incorporated to ensure comprehensive safety compliance across both short- and long-term horizons.
3) An adaptive shielding mechanism enforces state-wise instantaneous constraints by assessing action safety based on a pretrained partial transition model. Corrections are applied only when predicted violations exceed thresholds, ensuring strict constraint compliance with minimal interference to policy learning.

2. Model description

This study considers the NASA X-57 Mx aircraft retrofitted with a FC-battery hybrid propulsion system as the reference platform for FCHEA. The aircraft's maximum take-off weight remains the same as that of the original design, allowing the aerodynamic configuration and coefficients to be retained. The hybrid propulsion system is integrated and simulated using the open-source conceptual aircraft design platform SUAVE [39], which is extended for hybrid powertrain integration.

2.1. Aircraft and flight dynamics

A standard fixed-wing aerodynamic model is employed, where the lift force L balances the aircraft's weight and the drag force D determines propulsion demand. The forces are expressed as:

$$L = \frac{1}{2}\rho V^2 S C_L \tag{1}$$

$$D = \frac{1}{2}\rho V^2 S C_D \tag{2}$$

where C_L and C_D are the lift and drag coefficients; ρ is the atmospheric density at altitude h; S is the reference wing area; and V is the aircraft velocity.

To enable efficient flight trajectory prediction and energy management simulation, a longitudinal point-mass flight dynamics model is adopted. The aircraft state evolution is described in the Earth-fixed coordinate frame as shown in Fig. 1, with the equations of motion given by Ref. [40]:

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ h \\ V \\ \psi \\ \gamma \end{bmatrix} = \begin{bmatrix} V\cos(\psi)\cos(\gamma) \\ V\sin(\psi)\cos(\gamma) \\ V\sin(\gamma) \\ \frac{g}{W}[T\cos(\alpha) - D - W\sin(\gamma)] \\ \frac{g}{WV\cos(\gamma)}[T\sin(\alpha) + L]\sin(\mu) \\ \frac{g}{WV}[(T\sin(\alpha) + L)\cos(\mu) - W\cos(\gamma)] \end{bmatrix} \tag{3}$$

where x, y, and h denote the aircraft's position in inertial coordinates. T, D, L represent the aerodynamic forces of thrust, drag, and lift, respectively; α and γ present the angle of attack and the flight path; ψ and μ present the yaw heading angle and the roll bank angle; W is the aircraft weight. The propulsion power at each flight segment is computed by:

$$P_{drive} = \vec{T}\,\vec{V} = \frac{1}{2}C_D\rho S V^3 + m\frac{d}{dt}\left(\frac{1}{2}V^2\right) + mgV\sin\gamma \tag{4}$$

where $\vec{T}$ and $\vec{V}$ are the thrust and velocity vectors, m is the aircraft mass, g is the gravitational acceleration.

To support simulation and mission analysis, the open-source conceptual aircraft design platform SUAVE is extended for hybrid powertrain integration. In parallel, OpenVSP [41] is employed for geometric modelling and preliminary aerodynamic validation to improve the fidelity of aerodynamic estimations within SUAVE. Validation results in Fig. 2 demonstrate good agreement between the simulation and experimental wind tunnel data of X-57 Mx [42], with a deviation of less than ±4.67 %, confirming the validity of the aerodynamic model.

The topology of the proposed hybrid propulsion system is illustrated in Fig. 3. The FC system is connected to the DC bus through a boost

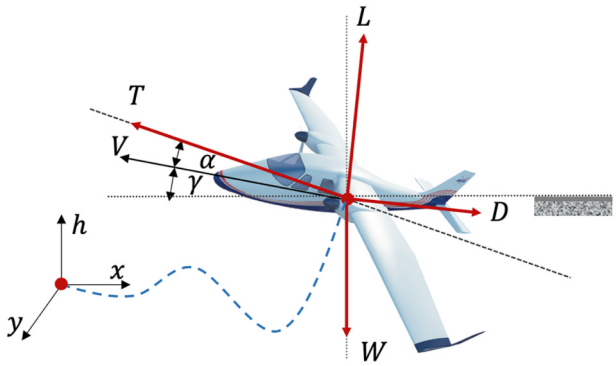


Fig. 1. Aircraft forces and position coordinate frame.

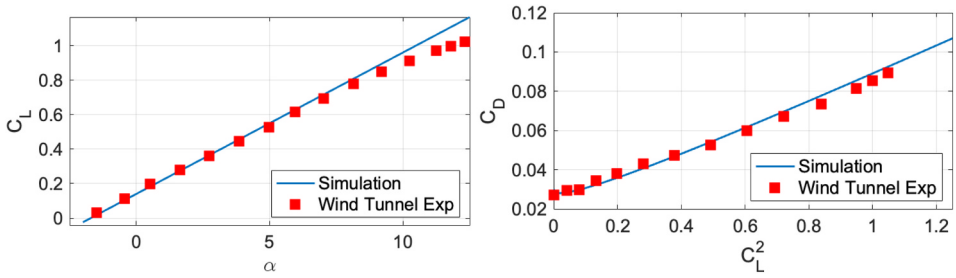


Fig. 2. Aerodynamic validation.

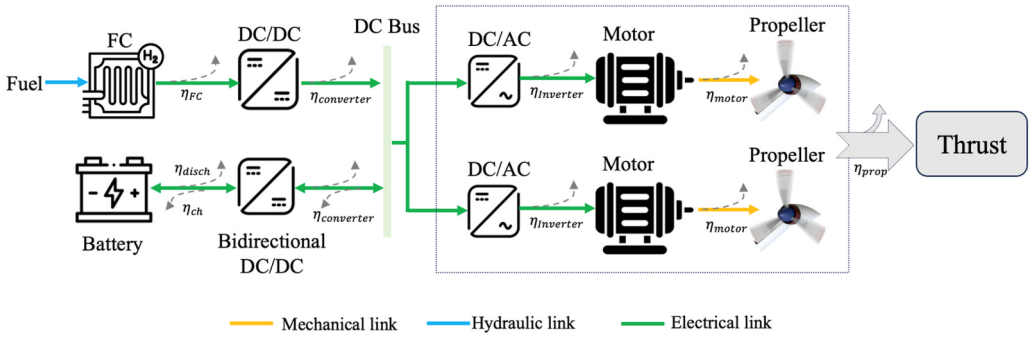


Fig. 3. Hybrid propulsion architecture.

converter, while the lithium-ion battery is connected via a bidirectional converter, enabling passively charging and discharging during the flight. Two electric propellers per propulsion segment are powered through inverters connected to the DC bus. This configuration supports dynamic power sharing between the FC and battery systems and forms the basis for the subsequent energy management optimization.

Additionally, the Dryden Wind Turbulence Model is employed to emulate atmospheric disturbances, where closed-loop interactions between wind disturbances and the aircraft's velocity and altitude profiles produce more realistic mission trajectories [43]. The wind components synthesized through shaping filters are as follows:

$$\frac{du(t)}{dt}+\frac{V}{L_u}u(t)=\frac{2\sigma_u\sqrt{V}}{\sqrt{\pi L_u}}w_u(t) \tag{5}$$

$$\frac{dw(t)}{dt}+\frac{V}{L_w}w(t)=\frac{2\sigma_w\sqrt{V}}{\sqrt{\pi L_w}}w_w(t) \tag{6}$$

where $u(t)$ and $w(t)$ are the turbulence velocity components, L_u and L_w are the turbulence scale lengths, σ_u and σ_w are the turbulence intensities, $w_u(t)$ and $w_w(t)$ are Gaussian white noise processes with zero mean for longitudinal and vertical directions.

2.2. Fuel cell modelling

A semi-empirical model [44] is employed to simulate the electrochemical behaviour of the FC and capture its internal voltage dynamics. The cell output voltage is determined based on the Nernst potential E_{Nernst} and associated polarization losses, including activation v_{act}, ohmic v_{ohmic}, and concentration v_{con} overpotentials, calculated as follows:

$$v_{FC}=nv_{cell} \tag{7}$$

$$v_{cell}=E_{Nernst}-v_{act}-v_{ohmic}-v_{con} \tag{8}$$

where n is the number of cells.

Hydrogen fuel consumption $\dot{m}_{H_2}$ is quantified from the instantaneous output power and FC efficiency:

$$\dot{m}_{H_2}=\frac{1}{LHV}\int_t^{t+\Delta T}\frac{P_{FC}(t)}{\eta_{FC}(t)}dt \tag{9}$$

where LHV represents the lower heating value of hydrogen, ΔT denotes the EMS control interval, P_{FC} and η_{FC} are the output power and operation efficiency of FC.

FC degradation remains a complex and active area of research due to the combined effects of electrochemical, thermal, and mechanical aging mechanisms. To balance model fidelity and computational efficiency, an empirical degradation model is adopted in place of detailed physical or data-driven approaches. This model captures condition-dependent degradation through four primary operational stressors: start-stop cycles, prolonged idle periods, large power fluctuations, and sustained high-power operation [45]. The cumulative degradation is represented as a weighted sum of independent contributions from each factor, with the state of degradation of FC (SoD_{FC}) defined as:

$$SoD_{FC}=d_{ss}n_{ss}+d_{idle}t_{idle}+d_vt_v+d_{high}t_{high} \tag{10}$$

where the coefficients and degradation rates are listed in Table 2.

Table 2
Parameters for FC degradation [45].

Coefficient	Definitions	Unit	Value
d_{ss}	Degradation per full start-stop	%/cycle	0.00196
n_{ss}	Number of full start-stop cycle	Count	–
d_{idle}	Degradation rate during idle operation (<5 % max power)	%/h	0.00126
t_{idle}	Idle duration	Hours	–
d_v	Degradation rate under load variation (>10 % of max power)	%/h	0.0000593
t_v	Duration of large load variation	Hours	–
d_{high}	Degradation rate under high load (>90 % max power)	%/h	0.00147
t_{high}	Duration of high load condition	Hours	–

2.3. Battery pack modelling

The battery pack is modelled using an equivalent circuit based on Thevenin's theorem. The relationship between internal resistance and open-circuit voltage as a function of SoC is illustrated in Fig. 4 [46]. The current output under a given power load and SoC- dependent internal resistance is given by:

$$I(t) = \pm \frac{v_{OC}(t) - \sqrt[2]{v_{OC}(t)^2 - 4R_{battery}(SoC)P_{battery}(t)}}{2\,R_{battery}(SoC)} \tag{11}$$

where I, $P_{battery}$, $R_{battery}$ andv_{OC} are current, output power, internal resistance and open circuit voltage of battery, respectively.

The SoC evolution is determined by the following integration over a control interval ΔT:

$$\Delta SoC(t) = \begin{cases} -\dfrac{1}{Q_{battery}} \dfrac{1}{\eta_{discha}} \displaystyle\int_t^{t+\Delta T} I(t)dt, I(t) > 0 \\ -\dfrac{1}{Q_{battery}} \eta_{cha} \displaystyle\int_t^{t+\Delta T} I(t)dt, I(t) < 0 \end{cases} \tag{12}$$

$$SoC(t+1) = SoC(t) + \Delta SoC(t) \tag{13}$$

where η_{discha} and η_{cha} represent battery discharging and charging efficiencies, $Q_{battery}$ represents energy storage capacity of battery.

A throughput-based capacity degradation model is employed to quantify the capacity loss of batteries, in which the degradation strongly depends on the C-rate. High C-rates enhance the dynamic responsiveness of batteries but simultaneously accelerate degradation due to increased electrochemical and mechanical stress [47]. The degradation behaviour is therefore described by an Arrhenius-type empirical model. The activation energy is expressed as a function of C-rate:

$$E_a = 31700 - 370.3\, c_{rate} \tag{14}$$

And for a given C-rate at a constant operating temperature T = 298.5 K in this study, the cumulative capacity loss as a function of ampere-hour throughput $A_h(t)$ is formulated as:

$$Q_{loss}(t) = d_{battery}(c_{rate})\, exp\left[-\frac{E_a(c_{rate})}{RT}\right] [A_h(t)]^z \tag{15}$$

where $d_{battery}(c_{rate})$ is the pre-exponential factor varying with c-rate referred to Table 3 [48], R is universal gas constant, 8.314 J/mol/K; z is the power-law factor equals to 0.55, A_h denotes the cumulative ampere-hour throughput.

2.4. Mission analysis and conceptual sizing

Given the weight sensitivity of aircraft, the energy onboard must match the mission energy demand. Excess energy onboard leads to unnecessary weight and drag penalties, which in turn increase fuel consumption and accelerate degradation of power components. Therefore, a discrete-time iterative algorithm is adopted to size the FC and battery subsystems under mission-dependent power demand. The FC serves as the primary power source and is sized to sustain cruise with maximum speed (172mph at 8000 feet), which defined by the X-57 design manual [49]. The battery is sized to provide additional power for transient conditions such as climb, acceleration, or load variation, thereby reducing stress on the FC. The process converges when the computed power profile matches the mission demand under mass and aerodynamic feedback, resulting in an energy-balanced propulsion system. The workflow of the hybrid-electric mission analysis is illustrated in Fig. 5, and the aircraft parameters are presented in Table 4.

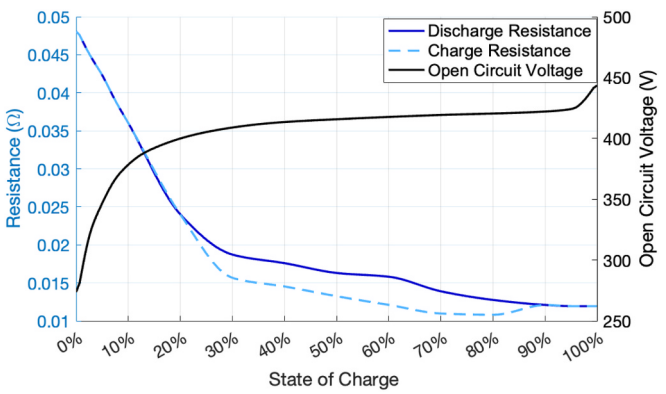


Fig. 4. Internal resistance and open circuit voltage of the battery.

Table 3
Dependence of battery degradation factor to c-rate [48].

c_{rate}	0.5	2	6
$d_{battery}$	31,630	21,681	12,934

3. Preliminaries and problem formulation

3.1. State-constrained Deep Deterministic Policy Gradient

In safety-critical decision-making problems, Constrained Markov Decision Processes (CMDP) provide a mathematical framework for optimizing sequential decisions under safety or resource limitations. A state-wise CMDP is defined as a tuple $(\mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R}, \delta, \mathscr{C})$, where S is the state space, $\mathscr{A}$ is the action space that an agent can take; $\mathscr{P} : \mathscr{S} \times \mathscr{A} \times \mathscr{S}$ is the state transition function; $\mathscr{R}$ is the immediate reward function; $\delta \in [0, 1)$ is a discount factor from first time-step onwards; and $\mathscr{C} = \{c_i\}$ is the set of constraint functions, where each c_i quantifies a specific cost associated with a state-action pair. The optimization objective is to find the policy π^* that maximizes the expected discounted total reward while satisfying constraints, as expressed below:

$$\pi^* = arg\, max_\pi\, J = \mathbb{E}\left[\sum_0^\infty \delta^t R(s_t, a_t)\right] s.t. c_t \le C_t, \forall i \tag{16}$$

To solve continuous control problems within the CMDP framework, Deep Deterministic Policy Gradient (DDPG) offers a strong foundation. It is a model-free, off-policy RL algorithm designed for continuous action spaces, which combines an actor-critic architecture with deterministic policy optimization. The actor network $\mu(s|\theta^\mu)$ maps states to actions, while the critic network $Q^\mu(s, a|\theta^Q)$ estimates the action-value function. The critic is trained using the temporal-difference loss:

$$L(\theta^Q) = \mathbb{E}_{(s,a,r,s')\sim M}\left[\left(Q(s, a|\theta^Q) - y_t\right)^2\right] \tag{17}$$

$$y_t = r + \delta Q'\left(s', \mu(s'|\theta^{\mu'})|\theta^{Q'}\right) \tag{18}$$

The Actor network updates its parameters to maximize the expected return based on the deterministic policy gradient:

$$\nabla_{\theta^\mu} J \approx \mathbb{E}_{s\sim M}\left[\nabla_a Q(s, a|\theta^Q)\big|_{a=\mu(s|\theta^\mu)} \nabla_{\theta^\mu}\, \mu(s|\theta^\mu)\right] \tag{19}$$

Target networks are softly updated to improve stability:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \tag{20}$$

For clarity, the key notations used in this section are summarized in Table 5.

3.2. Problem description and formulation

This study aims to develop an EMS that optimally allocates the power

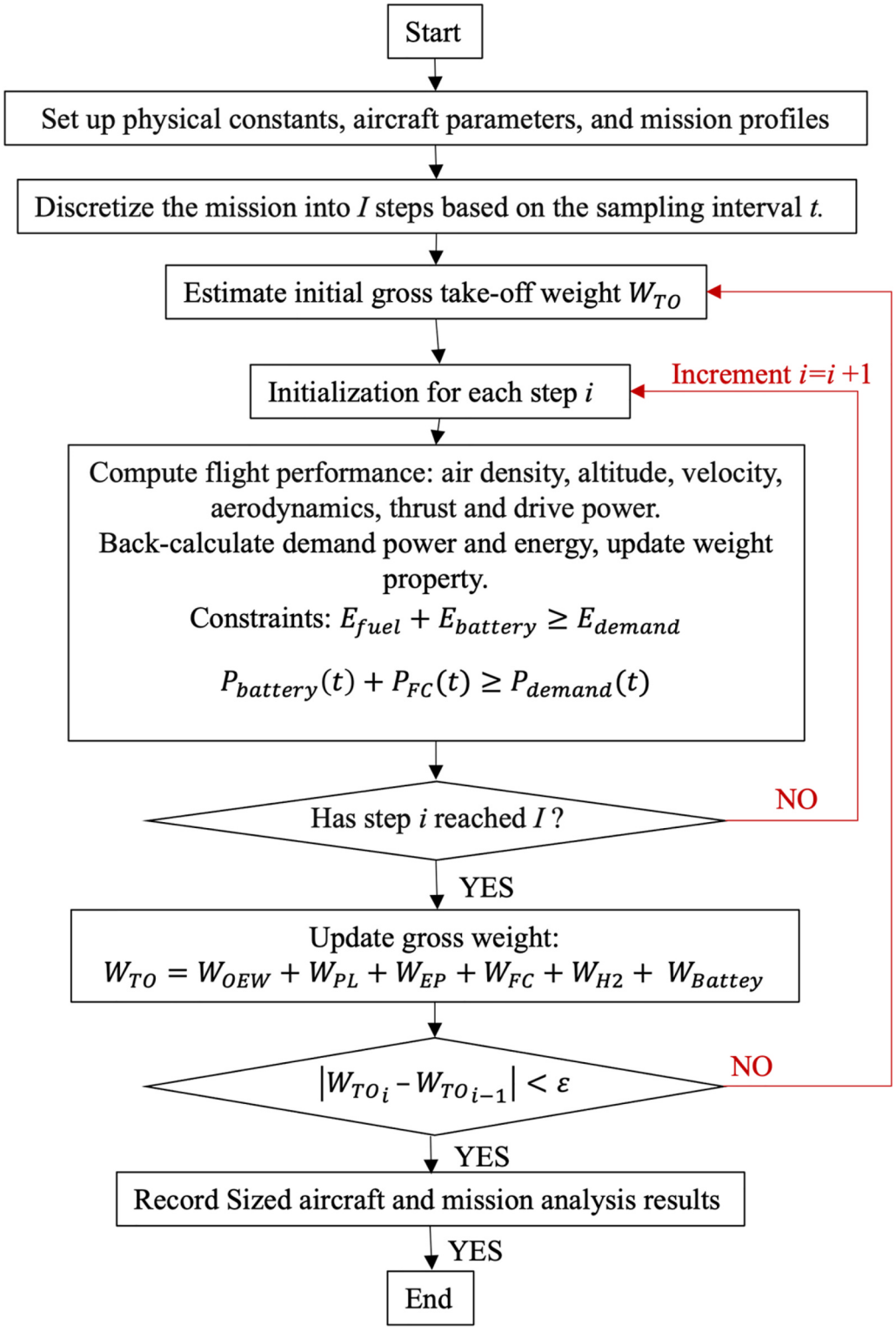


Fig. 5. Flow chart of hybrid electric aircraft mission analysis.

split between the FC and the battery to satisfy the time-varying propulsion power demands during flight. The EMS problem is formulated as a CMDP, aiming to minimize the total operational cost over the mission duration while satisfying multiple types of operational constraints.

The system state S_t represents the agent's observation at time t, comprising system health indicators and flight condition variables:

$$S_t = [SoD_{FC}(t), SoD_{Battery}(t), SoC(t), \alpha(t), \gamma(t), h(t), Ma(t), V(t), P_{demand}(t)] \quad (21)$$

where each component corresponds to FC degradation, battery degradation, state-of-charge, angle of attack, flight path angle, altitude, Mach number, airspeed, and propulsion power demand, respectively.

The action $a_t \in [0, 1]$ denotes the normalized power output of the FC,

Table 4
Specific parameters of FCHEA.

	Parameters	Values	Units
Aircraft	Aircraft Category	All-electric light aircraft	–
	Take off weight	920	kg
	Reference wing area	14.76 × 2	m^2
FC	Maximum power	150	kW
	FC rated voltage	460	V
	FC power density	2000	W/kg
	Power Ramp rate	45	kW/s
Battery	Battery rated capacity	25	Ah
	Battery rated voltage	460	V DC
	Battery energy density	1000	Wh/kg
	Maximum C-rate	6	h^{-1}

Table 5
Nomenclature.

Nomenclature	
$\mathscr{A}$	Action space
a_t	Action at time step t
$\mathscr{C}$	Set of constraint functions
c_t	Constraint function at time step t
$\mathbb{E}$	Expectation over the sampled batch of experience tuples
J	Expected cumulative reward under policy
$L(\theta^Q)$	Critic loss function
M	Replay memory buffer
$\mathscr{P}$	State-transition probability function
$\mathscr{R}$	Set of all immediate reward
R_t	Immediate reward received at time step t
$\mathscr{S}$	State space
s_t	State at time step t
s'	Next state
y_t	Target Q-value for critic update
τ	Target network soft update coefficient
$\mu(s\|\theta^\mu)$	Deterministic policy function (actor network) that maps state s to action, parameterized by θ^μ
δ	Discount factor
π	Deterministic policy
$Q^\mu(s, a\|\theta^Q)$	Predicted Q-value by the critic network for state–action pair (s, a), parameterized by θ^Q
θ^μ	Parameters of the actor network
$\theta^{\mu'}$	Parameters of the target actor
θ^Q	Parameters of the critic network
$\theta^{Q'}$	Parameters of the target critic
τ	Soft update rate for target networks

defined as:

$$a_t = \frac{P_{FC}(t)}{P_{FC_{max}}} \tag{22}$$

The objective is to derive a policy that maximizes cumulative reward, which in this context corresponds to minimizing the total operational cost. The stage-wise reward function R_t is formulated as:

$$-R_t = r_{fuel}(t) + r_{SoD_{FC}}(t) + r_{SoD_{battery}}(t) + C_{SoC}(t) \tag{23}$$

Each term in the cost function is defined as follows:

$$r_{fuel} = p_{fuel}\dot{m}_{H_2} \tag{24.1}$$

$$r_{SoD_{FC}} = \frac{p_{FC}SoD_{FC}}{EoL_{FC}} \tag{24.2}$$

$$r_{SoD_{battery}} = \frac{p_{battery}Q_{battery}SoD_{battery}}{EoL_{battery}} \tag{24.3}$$

$$C_{SoC}(t) = \omega\left|SoC - SoC_{ref}\right| \tag{24.4}$$

where r_{fuel} is the cost of hydrogen consumption; $r_{\mathrm{SoD_{FC}}}$ and $r_{\mathrm{SoD_{battery}}}$ are the depreciation cost of the FC and battery based on their degradation trajectories; C_{SoC} is a penalty for deviation from a reference SoC to maintain energy balance. The SoC bias weight ω is designed to be dimensionally compatible with fuel consumption and degradation. It penalizes deviations from the initial SoC to encourage charge recovery throughout the mission. In this study, ω was tuned via a discrete search to balance SoC recovery performance and training stability, with 0.1 selected as a suitable trade-off. Table 6 summarizes the parameters used in the reward function.

The action space and system states are constrained as follows:

$$P_{charge_{min}} \le P_{battery}(t) \le 0 \; if \; I(t) < 0 \tag{25.1}$$

$$0 \le P_{battery}(t) \le P_{discharge_{max}} \; if \; I(t) > 0 \tag{25.2}$$

$$P_{FCmin} \le P_{FC}(t) \le P_{FCmax} \tag{25.3}$$

$$0 \le SoD_{FC}(t) \le EoL_{FC} \tag{25.4}$$

$$0 \le SoD_{battery}(t) \le EoL_{battery} \tag{25.5}$$

$$SoC_{min} \le SoC(t) \le SoC_{max} \tag{25.6}$$

These constraints ensure the safe operation of the energy system components and the energy sufficiency across the mission profile.

4. Proposed safety mechanism

Due to the trial-and-error nature of RL-based energy management systems, explicit mechanisms are required to handle operational constraints and ensure safe policy learning and execution.

4.1. Constraint categories and enforcement

Safety constraints in this context can be categorized into three types: physical constraints, cumulative degradation constraints, and operational constraints. Each type demands a tailored enforcement strategy based on its structural properties. The classification and enforcement mechanisms of these constraints are summarized in Table 7.

Direct physical constraints, P_{FC} and $P_{battery}$, are enforced through hard-bounded logic within the environment. Any violations are either clipped or rejected to maintain physical feasibility. Cumulative constraints, SoD_{FC} and $SoD_{battery}$, are unidirectional and accumulate over time. These are not directly controllable in the short term but are penalized in the reward function to encourage long-term health-aware behaviour. Derived operational constraint, SoC, requires real-time regulation within a practical range [0.2,0.9] to avoid overcharging and deep discharging. This constraint is enforced using an adaptive shielding mechanism that evaluates potential violations at each decision step and applies minimal corrections when necessary.

4.2. Offline pretraining for constraint evaluation

To enable safe decision-making under constraints, a safety layer is constructed by learning an explicit approximation of the constraint transition dynamics. Specifically, a constraint model is trained to estimate the linearized sensitivity of constraint violation with respect to the

Table 6
Parameters for reward function.

Coefficient	Definitions	Value
p_{fuel}	Price of hydrogen fuel	£4/kg [50]
p_{FC}	Replacement cost of FC	£25,000 [51]
$p_{battery}$	Replacement cost of battery	£110/kWh [52]
EoL_{FC}	End-of-life degradation threshold for FC	10 %
$EoL_{battery}$	End-of-life degradation threshold for battery	20 %
SoC_{ref}	Reference SoC for battery management	60 %
ω	Weighting factor for SoC deviation	0.1

Table 7
Constraint typology and enforcement strategy.

Type		Variable	Description	Implementation Approach
Direct	Physical	$P_{battery}$, P_{FC}	Physically infeasible states	Clipped in the environment
Derived	Cumulative	SoD_{FC}, $SoD_{battery}$	Unidirectional and accumulative degradation; short-term violations are allowed	Penalized in the reward function
Derived	Operational	SoC	State-wise bounds for safe and efficiency operation	Corrected via adaptive shielding when violations predicted.

action. Instead of modelling entire trajectories, a single-step dynamics formulation is adopted to improve computational efficiency while maintaining sufficient accuracy for constraint evaluation. In this context, the next state SoC_{t+1} is determined by the current SoC_t and battery energy balance, which passively depends on P_{demand} and P_{FC}. Accordingly, the constraint-relevant state vector is defined as $s_t^c = \{SoC(t), P_{demand}, P_{FC}\}$. A pretraining dataset of 108,000 samples is collected through policy-agnostic uniform action sampling, ensuring broad coverage of the state–action space and generalization across diverse operating conditions throughout the mission profile. A feedforward neural network is employed to approximate the scalar constraint violation risk $c(s,a)$, using the following linear approximation:

$$\mathscr{T}(s)^\top = \frac{\partial c(s, a_{raw})}{\partial a_{raw}} \tag{26}$$

$$c(s, a^*) = \mathscr{T}(s)^\top (a_{raw} - a^*) + c(s, a_{raw}) \tag{27}$$

Here, $\mathscr{T}(s)$ is the gradient of the constraint with respect to action, predicted by the trained constraint model; a_{raw} denotes the raw action output by the original policy, a^* denotes the corrected action returned by the safe layer.

4.3. Adaptive shielding

To preserve policy performance while ensuring safety, a lightweight adaptive shielding mechanism is employed to apply minimal corrections only when activated. The shielding module computes a corrected action a^* by solving the following optimization problem:

$$\min_{a^*} \|a_{raw} - a^*\| = \delta a \tag{28.1}$$

$$\operatorname{argmin}_{a^*} \frac{1}{2}\|\delta a\|^2, s.t.\ c(s, a^*) \le C \tag{28.2}$$

Substituting the linearized constraint approximation $c(s, a^*)$, the constraint becomes:

$$\mathscr{T}(s)^\top (a_{raw} + \delta a) + c(s, a_{raw}) \le C \tag{29}$$

To derive the optimal correction δa, the Lagrangian formulation is constructed:

$$\mathscr{L}(\delta a, \lambda) = \frac{1}{2}\|\delta a\|^2 + \lambda(\mathscr{T}(s)^\top (a_{raw} + \delta a) + c(s, a_{raw}) - C) \tag{30}$$

The optimality conditions are characterized by satisfying the Karush-Kuhn-Tucker (KKT) conditions, as formulated:

$$\frac{\partial \mathscr{L}(\delta a, \lambda)}{\partial \delta a} = \delta a + \lambda\, \mathscr{T}(s) = 0 \tag{31}$$

Substituting back into the constraint yields the optimal Lagrange multiplier

$$\lambda^* = \frac{\mathscr{T}(s)^\top a_{raw} + c(s, a_{raw}) - C}{\mathscr{T}(s)^\top \mathscr{T}(s)} \tag{32}$$

The Lagrange multiplier λ^* represents the indicator of the Euclidean distance required to project the raw action back into the safe set. If Lagrange multiplier is non-negative, no constraint violation is predicted, and the raw action is applied directly. Otherwise, the raw action lies outside the safe set, a minimal correction is activated. Thus, the final corrected action is given by:

$$\lambda_{t+1} = max(0, \lambda^*) \tag{33}$$

$$a^* = \begin{cases} a_{raw} - \lambda_{t+1}\, \mathscr{T}(s), if\ \lambda > 0\ (violation) \\ a_{raw}, \lambda \le 0, otherwise\ (no\ violation) \end{cases} \tag{34}$$

In the proposed framework, the safety mechanism is externally integrated as a modular component, independent of the policy itself. The two-stage shielding process corresponds to section 4.2 and 4.3, as illustrated in Fig. 6. In contrast to conventional projection-based safety approaches that treat all actions as potentially unsafe and apply correction at every time step, the proposed state-wise adaptive shielding operates locally in a closed-loop manner. Real-time evaluation and correction are applied only when a constraint violation is predicted. The base policy is trained via DDPG, using a SumTree-structured prioritized experience replay [53] for efficient exploration. The shielding module operates externally during deployment, enabling safe control execution without interfering with policy training. The pseudo-code and framework structure are illustrated in Algorithm 1 and Fig. 7.

Algorithm 1 LC-EMS

Offline pretraining:
Construct the partial transition pretrain model $\mathscr{T}(s)$ to approximate the constraint dynamics using policy-agnostic uniformly sampled actions across complete missions.
Define the state-wise constraint model.
$\mathscr{T}(s)^\top = \frac{\partial c(s, a_{raw})}{\partial a_{raw}}$
RL Training:
Initialize *Critic* network $Q(s,a|\theta^Q)$ and *Actor* network $\mu(s|\theta^\mu)$.
Initialize memory buffer M with the capacity of $Memory_{max}$, incorporating a priority mechanism.
for episode = 1 to 400 **do**
  Initial noise for exploration N for exploration and Reset environment
  Obtain the initial state s
  for time step t = 1 to T **do**
    Select action $a_{raw,t} = \mu(s_t|\theta^\mu) + N_t$
    Evaluate safety constraints using adaptive shielding:
    If $\lambda_{t+1} = \lambda^*, a^* = a_{raw} - \lambda_{t+1}\, \mathscr{T}(s)$, **else** $a^* = a_{raw}$
    Execute safe action a^*, obtain reward r_t and next state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in reply buffer M with initial priority p.
      Sample minibatch (s_i, a_i, r_i, s_{i+1}) of size N from M with probability:
$P(i) = \frac{{p_i}^\alpha}{\sum_k {p_k}^\alpha}$
    Calculate importance-sampling weight: $w_i = \left(\frac{1}{N}\frac{1}{P(i)}\right)^\beta$
    Update *Critic* network by minimizing the loss L.
$L(\theta^Q) = \mathbb{E}_{s,a,r,s'\sim\mathscr{D}}\left[\left(Q(s,a|\theta^Q) - y_t\right)^2\right]$
    Update Actor network using policy gradient.
$\nabla_{\theta^\mu} J \approx \mathbb{E}_{s\sim\mathscr{D}}\left[\nabla_a Q(s,a|\theta^Q)\big|_{a=\mu(s|\theta^\mu)} \nabla_{\theta^\mu}\, \mu(s|\theta^\mu)\right]$
    Update the target networks:
$\theta^Q \leftarrow \tau\theta^Q + (1-\tau)\theta^Q$
$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$
  end for
end for

5. Results and discussion

5.1. Training configuration and learning process

The policy network and value network are updated via the DDPG algorithm, which interacts with the environment to collect experience

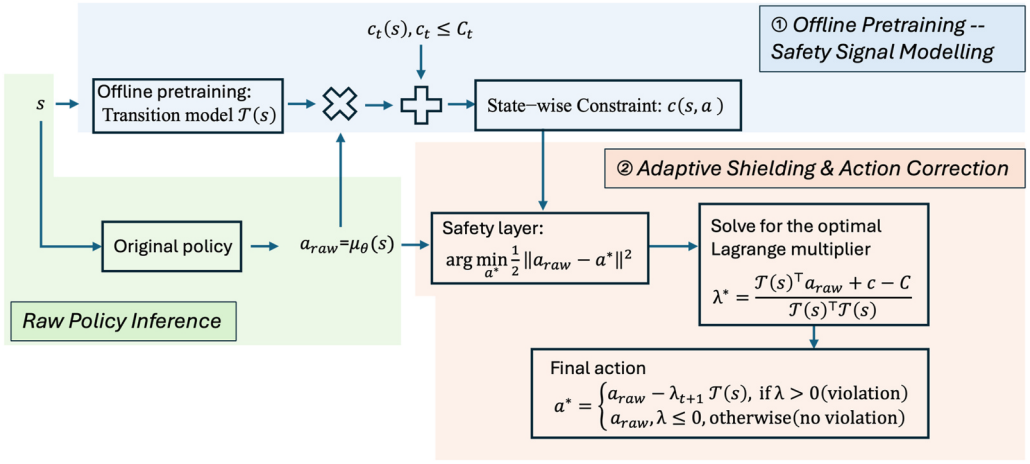


Fig. 6. Two-stage architecture of the proposed shielding mechanism.

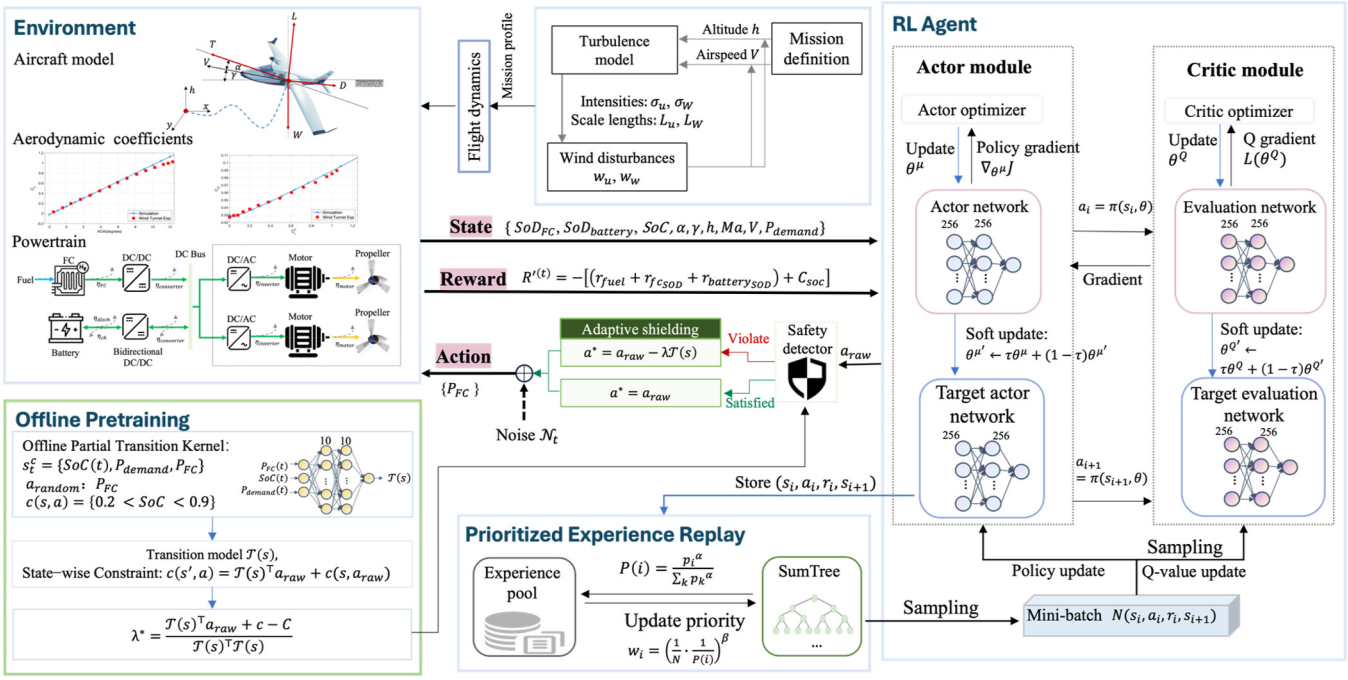


Fig. 7. Schematic of the LC-SEMS framework.

samples. A task-based evaluation protocol is adopted to assess policy generalization. To ensure the generalization capability of the proposed EMS across diverse mission scenarios, four representative flight missions are designed. These include one envelope-based mission defined by the Maxwell X-57 design manual [49], along with three mission profiles located within the operational boundary. A stochastic Dryden Wind Turbulence model with varied random seeds is introduced during training to enhance robustness and increase trajectory diversity. The detailed training mission profiles are illustrated in Fig. 8 and summarized in Table 8. The key hyperparameters of proposed LC-SEMS algorithm are summarized in Table 9.

To evaluate the adaptability and safety performance of the learned strategy, a validation mission composed of four sequential mission segments is constructed. This test mission combines multiple operating conditions within a single trajectory, simulating long-duration operation with cumulative component degradation. Between mission segments, hydrogen refuelling is assumed, while the battery is not recharged, introducing additional challenges to energy scheduling. For fair

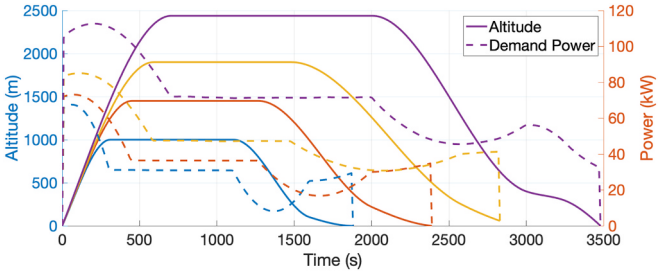


Fig. 8. Flight mission profile.

comparison among EMS strategies, the Dryden turbulence model is applied with a fixed random seed (42), ensuring consistent environmental disturbances across all algorithms during validation, as shown in Fig. 9.

The training performance is shown in Fig. 10, comparing discounted

Table 8
Typical flight mission definition.

Parameters	Mission 1	Mission 2	Mission 3	Envelope-based Mission
Cruise altitude (m)	1000	1450	1900	2438 (8000 ft)
Cruise speed (knots)	83	103	123	150
Flight range (nmi)	47.09	67.74	94.76	136.20
Duration (s)	1880	2390	2830	3480

Table 9
Key hyperparameters of proposed algorithm.

	Parameters description	Values
Pretraining	Neurons distribution of network	[10 10]
	Activation function	ReLU
	Learning rate	1e-3
	Number of pretraining data	108,000
Safe RL	Neurons distribution of Actor and Critic network	[256,256]
	Learning rate range of Actor network	5e-4
	Learning rate range of Critic network	5e-4
	Optimizer	Adam
	Capacity of experience replay buffer	100,000
	Number of training episodes	400
	Discounting factor γ	0.99
	Minibatch size	64
	Target update rate τ	0.005

reward curves of the baseline DDPG (B-DDPG) and the proposed LC-SEMS. B-DDPG exhibits significant oscillations during early episodes and converges after 175 episodes, whereas the LC-SEMS converges more rapidly and stabilizes after approximately 37 episodes. This improved convergence behaviour is attributed to the offline pretraining data and priority experience replay, which provides informative prior knowledge to guide early learning and reduces inefficient trial-and-error exploration.

5.2. Safety analysis of adaptive shielding

A comparative safety analysis is conducted across four flight missions to evaluate the safety performance of three representative EMS strategies: (1) B-DDPG, (2) reward shaping DDPG (RS-DDPG), and (3) the proposed LC-SEMS with adaptive safety shielding.

The B-DDPG agent optimizes purely for reward without any constraint enforcement. The RS-DDPG introduces a soft constraint mechanism by modifying the SOC bounds via a bound slack σ, forming an extended interval $(\underline{SOC}+\sigma, \overline{SOC}-\sigma)$.The corresponding shaping penalty function is defined as:

$$C_{soc} = \omega\left(max(0, \underline{SOC}+\sigma-SOC_i)^2\right) + \left(max\left(0, SOC_i-\overline{SOC}+\sigma\right)^2\right) \quad (35)$$

To quantify safety performance, a cumulative constraint violation metric V_c is defined:

$$V_c = \sum_{i=1}^{T} 1_{SOC_i \notin (\underline{SOC}, \overline{SOC})} [max(\underline{SOC}-SOC_i, 0) + max(SOC_i-\overline{SOC}, 0)] \quad (36)$$

To evaluate the robustness of constraint handling under varying mission conditions, each EMS strategy was evaluated across four flight missions using 8 random seeds, and the cumulative SOC constraint violations were visualized in Fig. 11. In Mission 1 and Mission 2, where power and energy demands are relatively low (71.65 kW/16.92 kWh and 76.83 kW/25.04 kWh, respectively), all strategies demonstrate acceptable constraint adherence. However, B-DDPG exhibits small violations with a mean of 0.02×10^3 ($\pm 0.04 \times 10^3$) and 0.31×10^3 ($\pm 0.25 \times 10^3$), respectively. In Mission 3 and the Envelope Mission, where the peak power and energy demands are significantly higher (88.97 kW/38.67 kWh and 117.56 kW/68.26 kWh), both B-DDPG and RS-DDPG struggle to meet the SOC constraints. B-DDPG incurs large violations with a mean of 2.83×10^3 ($\pm 1.61 \times 10^3$) in Mission 3 and 7.06×10^3 ($\pm 2.38 \times 10^3$) in the Envelope Mission. RS-DDPG shows moderate improvement (0.23×10^3 $\pm 0.26 \times 10^3$ and 1.38×10^3 $\pm 0.55 \times 10^3$, respectively) but still fails to eliminate violations under high-demand conditions. In contrast, the proposed LC-SEMS strategy maintains zero constraint violations across all missions, confirming the effectiveness of the adaptive shielding mechanism in proactively preventing unsafe actions and ensuring safety under diverse mission profiles. The pretraining strategy effectively reduces inefficient trial-and-error exploration during the early stages of training and fundamentally addresses the problem of early episode termination caused by constraint violation, thereby improving learning stability and accelerating convergence.

5.3. EMS performance evaluation

The power management performance is compared against three baseline methods: a rule-based FL controller, a DP solution as a global offline optimal benchmark, and a classical model-free RL-based Deep Q-

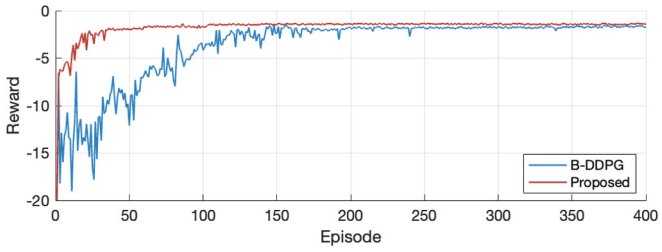


Fig. 10. Discounted reward of RL-based EMS.

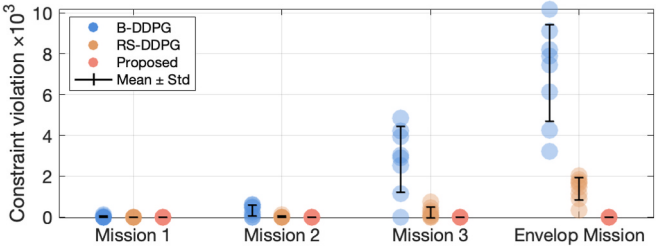


Fig. 11. Cumulative constraint violations of RL-based EMS.

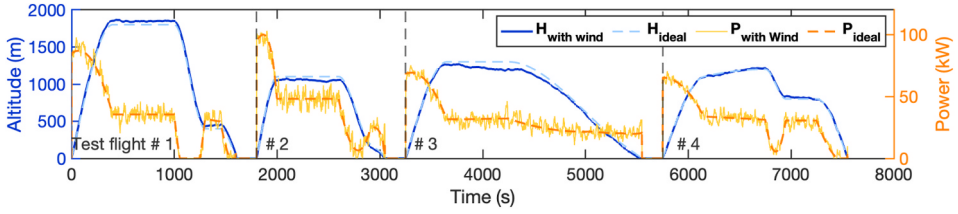


Fig. 9. Sequential Validation mission profile.

network (DQN) strategy.

The power distribution and SOC trajectories of the four EMS strategies are shown in Fig. 12. The rule-based FL controller, relying solely on current-state heuristics without temporal foresight or optimization, fails to achieve global planning. This results in sharp FC power oscillations under dynamic loads, for example, during the cruise phase of Mission 2 (2000–2550 s), severe FC output fluctuations occur alongside an SoC drop below 3.5 %, leading to increased hydrogen consumption and accelerated degradation. The DP strategy, functioning as a fully offline global optimization method, performs backward iteration over the entire state-action space with full access to mission information, generating smooth and demand-aligned power output and stable SOC trajectories. The DQN-based controller, trained online without mission priors, shows high-frequency power fluctuations, indicating poor generalization and suboptimal policy learning in the absence of mission-specific prior knowledge. The proposed LC-SEMS strategy integrates offline pretraining, SoC-aware safety estimation, and adaptive shielding to correct unsafe actions before execution. It achieves smoother control, with a FC power variance is 733.64 kW^2, representing an 8.21 % reduction compared to DQN. Despite minor delays during mission transitions, LC-SEMS demonstrates stable power allocation and approximates DP-level coordination under partial observability, while remaining applicable in real-time.

This proposed strategy focuses on the high-level strategy-making framework, assuming that the low-level controller is well designed to execute commands issued by the high-level decisions. The proposed framework achieves low latency through closed-form computation. On an Apple M2 Pro chip with 16 GB unified memory, the measured decision time is 0.21 s, which is well within a single-step interval and highlights its potential for real-time deployment in aviation energy management systems.

Fig. 13 compares SoC trajectories across all EMS strategies. While all end with SoC within acceptable bounds, the trajectory patterns vary significantly. The FL controller shows the largest SOC fluctuations and consistently lower SOC values, reflecting its reliance on handcrafted rules and limited adaptability to diverse scenarios. DP achieves the most stable trajectory due to complete mission foresight. In contrast, DQN exhibits mid-mission deviations caused by policy instability and limited coordination, leading to inefficient energy buffering. In comparison, the proposed LC-SEMS strategy achieves a more regulated SOC trajectory by leveraging current-state estimation and experience replay, with reduced deviation relative to DQN, indicating improved adaptability and control robustness in partially observable and dynamic environments.

The comprehensive cost analysis of the four EMS strategies is summarized in Table 10 and Fig. 14, where the total cost is composed of fuel consumption and the depreciation cost of both the FC and battery, calculated based on unit price and observed health loss. The DP strategy achieves the lowest total cost at £39.63 and serves as the benchmark. The proposed LC-SEMS strategy follows closely at £41.37, corresponding to a 95.79 % cost-saving rate relative to DP. From the perspective of component aging trade-off, due to the significant price gap between FC and battery, all three strategies adopt a similar control preference by shifting dynamic load to the battery to mitigate FC degradation. Although the battery experiences 5–10 times more physical degradation than the FC (e.g., under the proposed strategy: 0.4821 % vs. 0.0432 %), this trade-off achieves economic balance due to the battery's substantially lower replacement cost. DQN strategy leads to a FC depreciation cost of £19.59, while LC-SEMS reduces it to £13.22, accounting for the majority of the 15.86 % total cost reduction. Motivated by the observed cost differences in Fig. 14, where FC depreciation is identified as the primary driver of cost variation across strategies, Fig. 15 provides a further breakdown of FC degradation. The analysis reveals that output variant load is the dominant contributor, reaching 75.19 $\times 10^{-5}$ under DQN, whereas the proposed strategy suppresses it to 40.71 $\times 10^{-5}$, closely matching DP 44.12 $\times 10^{-5}$. These results confirm that minimizing power instability is critical to mitigating FC aging. This indicates that, despite being derived from a black-box RL-based model, the proposed strategy exhibits behaviour that can be interpreted in terms of underlying physical principles.

6. Conclusion

Given the current limitations of battery technologies, FCHEA are regarded as a promising transitional solution toward aviation decarbonization. This study presents a LC-SEMS for FCHEA, targeting long-term cost minimization and state-wise safety enforcement under

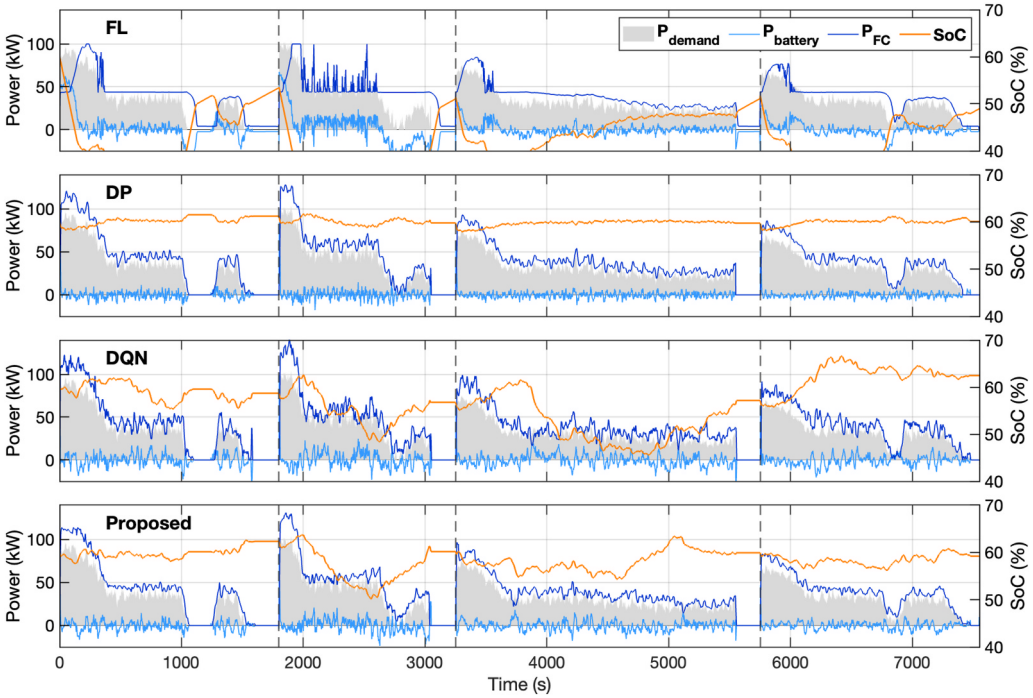


Fig. 12. Power distribution and SoC trajectory.

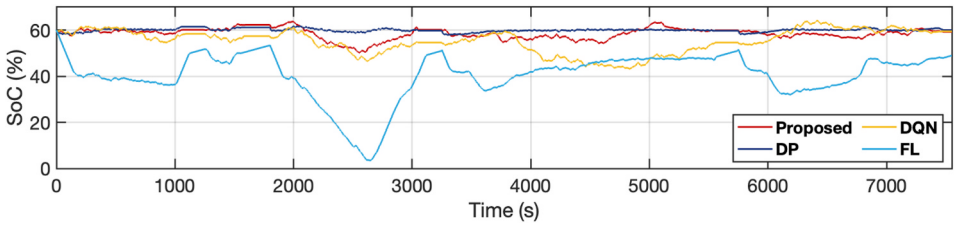


Fig. 13. SOC trajectories comparison.

Table 10
Performance summary of the four EMS strategies.

	FL	DP	DQN	Proposed
Fuel consumption (g)	4390.67	4310.76	4343.45	4326.03
FC SoD (%)	99.8896	99.9531	99.9216	99.9568
Battery SoD (%)	99.6072	99.6123	99.6012	99.5179
Depreciation (£)	38.41	22.39	30.56	24.07
Overall cost (£)	55.97	39.63	47.93	41.37
Cost-saving rate (%)	70.80	100	82.68	95.79

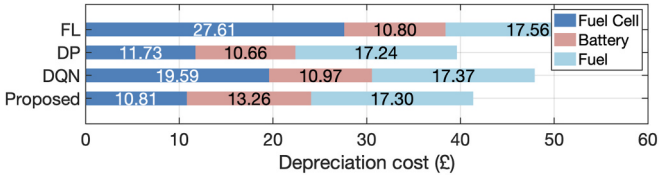


Fig. 14. Cost composition comparison.

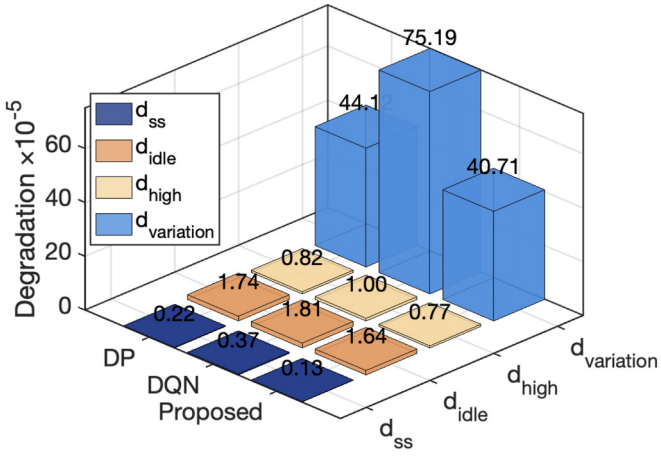


Fig. 15. FC degradation breakdown.

dynamic flight scenarios.

Firstly, a physics-informed RL environment was developed, integrating propulsion, aerodynamics, electric powertrain characteristics and flight dynamic to enable realistic and longevity-conscious policy training. Then, based on this environment, the energy management problem was formulated to explicitly optimize fuel consumption, battery SoC regulation, and component health under state-wise constraints. The adaptive shielding mechanism was then integrated into the EMS framework. It operates only when constraint-violating actions are predicted, enabling real-time action correction via a closed-form solution with negligible computational cost. It ensures state-wise constraint satisfaction through minimal action corrections, calculated in closed form with minimal computational cost, while leaving the training process fully unconstrained to enable effective exploration.

The proposed LC-SEMS was validated across multiple flight profiles with varying cruise altitudes and speeds under simulated wind disturbances. Results demonstrate that FC degradation is highly sensitive to power fluctuations. By enhancing load transition robustness, the LC-SEMS reduces FC depreciation by 44.82 % compared to the RL baseline. Overall, the strategy achieves a 95.79 % cost-saving rate while maintaining a performance gap of only 4.21 % relative to the globally optimal DP benchmark. These results suggest that the LC-SEMS offers a safe, adaptive, and cost-effective energy management solution for the next generation of hybrid electric aviation systems.

Despite these promising results, some limitations remain. While this approach is effective for proof-of-concept analysis, hardware-in-the-loop experiments are needed to support practical deployment. Future work will incorporate FC and battery testbeds to evaluate system performance under real-time conditions and introduce load fluctuation uncertainties to enhance the robustness and reliability of the EMS in real-world missions.

CRediT authorship contribution statement

Yajing Xiao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jinning Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Harold S. Ruiz:** Writing – review & editing, Supervision. **Ioannis Roumeliotis:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Xin Zhang:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Radomsky L, Keilmann R, Ferch D, Mallwitz R. Challenges and opportunities in power electronics design for all-and hybrid-electric aircraft: a qualitative review and outlook. CEAS Aeronaut J 2024;15(4):751–64.
[2] Hwang J, Maharjan K, Cho H. A review of hydrogen utilization in power generation and transportation sectors: achievements and future challenges. Int J Hydrogen Energy 2023;48(74):28629–48.
[3] Avogadro N, Redondi R. Demystifying electric aircraft's role in aviation decarbonization: are first-generation electric aircraft cost-effective? Transport Res Transport Environ 2024;130:104191.
[4] Xu L, Huangfu Y, Ma R, Xie R, Song Z, Zhao D, et al. A comprehensive review on fuel cell UAV key technologies: propulsion system, management strategy, and design procedure. IEEE Trans Transport Electrification 2022;8(4):4118–39.

[5] Gao Q, Lei T, Yao W, Zhang X, Zhang X. A health-aware energy management strategy for fuel cell hybrid electric UAVs based on safe reinforcement learning. Energy 2023;283:129092.
[6] Cardone L, Petrone G, De Rosa S, Franco F, Greco C. Review of the recent developments about the hybrid propelled aircraft. Aerotecnica Missili & Spazio 2024;103(1):17–37.
[7] Fernandez AM, Kandidayeni M, Boulon L, Chaoui H. An adaptive state machine based energy management strategy for a multi-stack fuel cell hybrid electric vehicle. IEEE Trans Veh Technol 2019;69(1):220–34.
[8] Maghfiroh H, Wahyunggoro O, Cahyadi AI. Energy management in hybrid electric and hybrid energy storage system vehicles: a fuzzy logic controller review. IEEE Access 2024;12:56097–109.
[9] Sidharthan Panaparambil V, Kashyap Y, Vijay Castelino R. A review on hybrid source energy management strategies for electric vehicle. Int J Energy Res 2021;45(14):19819–50.
[10] Zhang H, Li Q, Wang H, Li Q, Qin G, Wu Q. A review of energy management optimization based on the equivalent consumption minimization strategy for fuel cell hybrid power systems. Fuel Cells 2022;22(4):116–30.
[11] Tian W, Liu L, Zhang X, Shao J, Ge J. A coordinated optimization method of energy management and trajectory optimization for hybrid electric UAVs with PV/Fuel cell/battery. Int J Hydrogen Energy 2024;50:1110–21.
[12] Tang W, Wang Y, Jiao X, Ren L. Hierarchical energy management strategy based on adaptive dynamic programming for hybrid electric vehicles in car-following scenarios. Energy 2023;265:126264.
[13] Saiteja P, Ashok B. Critical review on structural architecture, energy control strategies and development process towards optimal energy management in hybrid vehicles. Renew Sustain Energy Rev 2022;157:112038.
[14] He H, Meng X, Wang Y, Khajepour A, An X, Wang R, et al. Deep reinforcement learning based energy management strategies for electrified vehicles: recent advances and perspectives. Renew Sustain Energy Rev 2024;192:114248.
[15] Zhang Y, Ma R, Zhao D, Huangfu Y, Liu W. A novel energy management strategy based on dual reward function Q-learning for fuel cell hybrid electric vehicle. IEEE Trans Ind Electron 2021;69(2):1537–47.
[16] Ghaderi R, Kandidayeni M, Boulon L, Trovão JP. Q-learning based energy management strategy for a hybrid multi-stack fuel cell system considering degradation. Energy Convers Manag 2023;293:117524.
[17] Zhang J, Tao J, Hu Y, Ma L. An energy management strategy based on DDPG with improved exploration for battery/supercapacitor hybrid electric vehicle. IEEE Trans Intell Transport Syst 2023;25(5):3999–4008.
[18] Mazzi Y, Sassi HB, Errahimi F, Es-Sbai N. A soft actor-critic reinforcement learning framework for optimal energy management in electric vehicles with hybrid storage. J Energy Storage 2024;99:113344.
[19] Hu K, Li M, Song Z, Xu K, Xia Q, Sun N, et al. A review of research on reinforcement learning algorithms for multi-agents. Neurocomputing 2024:128068.
[20] Wachi A, Shen X, Sui Y. A survey of constraint formulations in safe reinforcement learning. Proc 33rd Int Joint Conf Artif Intell 2024:8262–71.
[21] Viswanadhapalli JK, Elumalai VK, Shah S, Mahajan D. Deep reinforcement learning with reward shaping for tracking control and vibration suppression of flexible link manipulator. Appl Soft Comput 2024;152:110756.
[22] Zhang H, Peng J, Tan H, Dong H, Ding F. A deep reinforcement learning-based energy management framework with Lagrangian relaxation for plug-in hybrid electric vehicle. IEEE Trans Transport Electrification 2020;7(3):1146–60.
[23] Huang R, He H, Zhao X, Wang Y, Li M. Battery health-aware and naturalistic data-driven energy management for hybrid electric bus based on TD3 deep reinforcement learning algorithm. Appl Energy 2022;321:119353.
[24] Zhang L, Zhang R, Wu T, Weng R, Han M, Zhao Y. Safe reinforcement learning with stability guarantee for motion planning of autonomous vehicles. IEEE Transact Neural Networks Learn Syst 2021;32(12):5435–44.
[25] Zhang M, Guo G, Magnússon S, Pilawa-Podgurski RC, Xu Q. Data driven decentralized control of inverter based renewable energy sources using safe guaranteed multi-agent deep reinforcement learning. IEEE Trans Sustain Energy 2023;15(2):1288–99.
[26] Lin J-L, Hung W, Yang S-H, Hsieh P-C, Liu X. Escaping from zero gradient: revisiting action-constrained reinforcement learning via frank-wolfe policy optimization. Conference escaping from zero gradient: revisiting action-constrained reinforcement learning via frank-wolfe policy optimization. PMLR, p. 397-407.
[27] Hao G, Li Y, Li Y, Jiang L, Zeng Z. Lyapunov-based safe reinforcement learning for microgrid Energy Management. IEEE Transact Neural Networks Learn Syst 2024;35(12):1234–45.
[28] Yu P, Zhang H, Song Y, Wang Z, Dong H, Ji L. Safe reinforcement learning for power system control: a review. Renew Sustain Energy Rev 2025;223:116022.
[29] Ruan J, Liang G, Zhao J, Zhao H, Qiu J, Wen F, et al. Deep learning for cybersecurity in smart grids: review and perspectives. Energy Convers Economics 2023;4(4):233–51.
[30] Gu S, Yang L, Du Y, Chen G, Walter F, Wang J, et al. A review of safe reinforcement learning: methods, theories and applications. IEEE Trans Pattern Anal Mach Intell 2024;46(12):11216–35.
[31] Yadav S, Kumar P, Kumar A. Hybrid renewable energy systems design and techno-economic analysis for isolated rural microgrid using HOMER. Energy 2025:136442.
[32] Abdelsattar M, Ismeil MA, Aly MM, Abu-Elwfa SS. Energy management of microgrid with renewable energy sources: a case study in Hurghada Egypt. IEEE Access 2024;12:19500–9.
[33] Zhou B, Wang D, Yang Z, Dong Y, An Y, Peng W. Adaptive equivalent consumption minimization strategy with battery aging considerations for parallel hybrid electric vehicle. Conference adaptive equivalent consumption minimization strategy with battery aging considerations for parallel hybrid electric vehicle. IEEE, p. 102-107.
[34] Ye Y, Wang H, Xu B, Zhang J. An imitation learning-based energy management strategy for electric vehicles considering battery aging. Energy 2023;283:128537.
[35] Ren X, Ye J, Xie L, Lin X. Battery longevity-conscious energy management predictive control strategy optimized by using deep reinforcement learning algorithm for a fuel cell hybrid electric vehicle. Energy 2024;286:129344.
[36] Ren J, Ma J, Wang H, Yu T, Wang K. A comprehensive review on research methods for lithium-ion battery of state of health estimation and end of life prediction: methods, properties, and prospects. Protect Control Mod Power Syst 2024;10(3):146–65.
[37] Lee J, Ko S, Kim B-K, Ryu J-H, Baek JD, Kang S-W. Empirical lifetime prediction through deterioration evaluation of high-power PEMFC for railway vehicle applications. Int J Hydrogen Energy 2024;71:972–81.
[38] Szrama S, Lodygowski T. Aircraft engine remaining useful life prediction using neural networks and real-life engine operational data. Adv Eng Software 2024;192:103645.
[39] Silva GR, Päßler J, Zwanepol J, Alberts E, Tarifa SLT, Gerostathopoulos I, et al. SUAVE: an exemplar for self-adaptive underwater vehicles. Conference SUAVE: an exemplar for self-adaptive underwater vehicles. IEEE, p. 181-187.
[40] Hull DG. Fundamentals of airplane flight mechanics. Springer; 2007.
[41] Kim J, Tsuchiya T. OPENVSP based aerodynamic design optimization tool building method and its application to tailless UAV. Conference OPENVSP based aerodynamic design optimization tool building method and its application to tailless UAV.
[42] Botero EM, Clarke MA, Erhard RM, Smart JT, Alonso JJ, Blaufox A. Aerodynamic verification and validation of SUAVE. Conference Aerodynamic verification and validation of SUAVE. 1929.
[43] Burgstaller I, Galffy A. A review of counter turbulence strategies, with a focus on gust load alleviation technology in manned aircraft. Conference A review of counter turbulence strategies, with a focus on gust load alleviation technology in manned aircraft. p. 4490.
[44] Bai F, Quan H-B, Yin R-J, Zhang Z, Jin S-Q, He P, et al. Three-dimensional multi-field digital twin technology for proton exchange membrane fuel cells. Appl Energy 2022;324:119763.
[45] Pei P, Chang Q, Tang T. A quick evaluating method for automotive fuel cell lifetime. Int J Hydrogen Energy 2008;33(14):3829–36.
[46] He L, Chen F, Tian P, Gou H. An improved energy management strategy for hybrid electric powered aircraft based on deep reinforcement learning. Aero Sci Technol 2024;149:109137.
[47] Rauf H, Khalid M, Arshad N. Machine learning in state of health and remaining useful life estimation: theoretical and technological development in battery degradation modelling. Renew Sustain Energy Rev 2022;156:111903.
[48] Li Y, Wang F, Tang X, Lin X, Liu C, Hu X. Real-time multiobjective energy management for electrified powertrains: a convex optimization-driven predictive approach. IEEE Trans Transport Electrification 2022;8(3):3139–50.
[49] Harris K. X-57 cruise motor and high-lift motor mission profile power analysis. 2021.
[50] Lane B, Reed J, Shaffer B, Samuelsen S. Forecasting renewable hydrogen production technology shares under cost uncertainty. Int J Hydrogen Energy 2021;46(54):27293–306.
[51] Halder P, Babaie M, Salek F, Shah K, Stevanovic S, Bodisco TA, et al. Performance, emissions and economic analyses of hydrogen fuel cell vehicles. Renew Sustain Energy Rev 2024;199:114543.
[52] Mauler L, Duffner F, Zeier WG, Leker J. Battery cost forecasting: a review of methods and results with an outlook to 2050. Energy Environ Sci 2021;14(9):4712–39.
[53] Wang C, Zhang J, Wang A, Wang Z, Yang N, Zhao Z, et al. Prioritized sum-tree experience replay TD3 DRL-Based online energy management of a residential microgrid. Appl Energy 2024;368:123471.



https://dspace.org/

Safe reinforcement learning-based energy management for fuel cell hybrid electric aircraft with longevity considerations

Xiao, Yajing; Zhang, Jinning; Ruiz, Harold S.; Roumeliotis, Ioannis; Zhang, Xin

2025-11-30
Elsevier
https://doi.org/10.1016/j.energy.2025.138782
Xiao Y, Zhang J, Ruiz HS, et al., (2025) Safe reinforcement learning-based energy management for fuel cell hybrid electric aircraft with longevity considerations. Energy, Volume 338, November 2025, Article number 138782

Downloaded from a DSpace repository. https://dspace.org/