Cost-effective reinforcement learning energy management for plug-in hybrid fuel cell and battery ships

Peng Wu*, Julius Partridge, Richard Bucknall

Marine Research Group, Department of Mechanical Engineering, University College London, London WC1E 7JE, UK

Abstract

Hybrid fuel cell and battery propulsion systems have the potential to offer improved emission performance for coastal ships with access to H_2 replenishment and battery charging infrastructures in ports. However, such systems could be constrained by high power source degradation and energy costs. Cost-effective energy management strategies are essential for such hybrid systems to mitigate the high costs. This article presents a Double Q reinforcement learning based energy management system for such systems to achieve near-optimal average voyage cost. The Double Q agent is trained using stochastic power profiles collected from continuous monitoring of a passenger ferry, using a plug-in hybrid fuel cell and battery propulsion system model. The energy management strategies generated by the agent were validated using another test dataset collected over a different period. The proposed methodology provides a novel approach to optimal use hybrid fuel cell and battery propulsion systems for ships. The results show that without prior knowledge of future power demands, the strategies can achieve near-optimal cost performance (96.9%) compared to those derived from using dynamic programming with the equivalent state space resolution.

Keywords: Coastal ferry, Hybrid fuel cell and battery, Continuous monitoring, Energy management system, Reinforcement learning

*Corresponding author

Email address: peng.wu.14@ucl.ac.uk (Peng Wu)

Nomenclature

Symbol	Description	Symbol	Description
Roman symbols		Δt	Time step, s
a	Action	δ	Degradation function
B	Battery capacity, kW h	ε	ε-greedy policy parameter
C	Battery C-rate	η	Efficiency
C_1	Battery cell model capacitance, F	π	Policy
i	Battery cell model current, A	σ	Price
$\dot{m}$	Mass flow rate, $\mathrm{kg\,s^{-1}}$	υ	GWP
N	Episode number	ψ	Fuel cell specific H_2 consumption
P	Power in system model, kW	**Subscripts**	
P	Probability in reinforcement learning	bat	Battery
Q	Action value function	fc	Fuel cell
R	Battery model resistance, Ω	t	$t-th$ time step
r	Reward	e	Electricity
s	Current state index	**Abbreviations**	
$s^{'}$	Next state index	AC	Alternating current
S	State parameter vector	DC	Direct current
spA	Shore power availability	DDP	Deterministic dynamic programming
SOC	Battery state of charge	EMS	Energy management system
x	Fuel cell state	GHG	Greenhouse gas
Greek symbols		GWP	Global warming potential
α	Learning rate	MDP	Markov Decision Process
γ	Discount factor	PEMFC	Proton exchange membrane fuel cell
$\Delta\varepsilon$	ε decay rate	RL	Reinforcement learning
$\Delta\alpha$	Learning rate decay rate	SOC	State of charge

1. Introduction

The hybridisation of Proton Exchange Membrane Fuel Cells (PEMFC) [1] with lithium batteries recharged from shore-generated electrical power [2] may potentially offer desirable emission performance for coastal ships with access to H_2 replenishing and battery charging infrastructures [3]. By utilising both H_2 and shore-generated electrical power, the life-cycle emission performance could be reduced significantly [3]. However, such hybrid systems are constrained by high degradation of the power sources and energy costs [4]. It is necessary to improve the operational cost-effectiveness of such hybrid systems.

In a hybrid propulsion system, an effective Energy Management System (EMS) is crucial to managing power utilisation from the multiple power sources. The EMS determines output actions of the hybrid propulsion system under certain operating conditions. However, it is challenging to develop effective EMS for hybrid systems since they need to be applied to load power profiles that have not been detailed in advance of the application. For a hybrid system with an energy storage device the problem, in essence, is that of sequential-decision making, i.e. what actions or controls should be taken over the power cycles to achieve optimal objectives (e.g. minimum operational costs, minimum emissions or a weighted balance). For hybrid road vehicles, standard driving cycles can be used to develop such EMS. However, for hybrid ships, such standard cycles do not exist and actual power demands over a series of voyages may vary significantly due to factors such as sea states, weather and cargo loading conditions. In recent years, continuous monitoring of power demands over the long term provides a potential new approach to develop effective EMS for such vessels [5].

The research of EMS for hybrid propulsion systems is primarily driven by road vehicle applications. Sulaiman et al. [6] provided a comprehensive review of the main EMS categories for hybrid fuel cell road vehicles. Their review indicates that energy management strategies such as rule-based, fuzzy logic, Equivalent Consumption Minimisation Strategy and wavelet-based load sharing are the main EMS streams for hybrid fuel cell road vehicles. Wang et al. [7] developed a rule-based on-line energy management strategy for hybrid fuel cell and ultracapacitor systems. Although near-optimal fuel economy has been achieved by their rule-based strategy in the two load profiles under investigation, the performance of the strategy in other profiles which can vary significantly is not clear. Wang et al. [8] proposed an energy management strategy based on velocity prediction for three typical non-plug in hybrid electric propulsion structures. An urban driving cycle was used to calculate the state transition probabilities; subsequently, dynamic programming was employed to generate the strategy. Recently, machine learning EMS has been applied to road vehicles. Muñoz et al. [9] presented a neural network EMS for hybrid fuel cell and battery road vehicles using supervised learning. With a target EMS for specific driving cycles generated by optimisation approaches, the neural network was subsequently trained to achieve performance similar to that achieved by the target EMS. The actual performance of such EMS when applied to unknown driving cycles is not clear. Fletcher et al. [10] adopted stochastic dynamic programming to generate optimal EMS for a hybrid fuel cell/battery road vehicle, accounting for the fuel cell degradation characteristics generalised from experimental results. Their EMS was able to reduce the cost by 12.3% due to prolonged fuel cell lifetime. However, the accuracy of stochastic dynamic programming is limited by its resolution due to 'the curse of dimensionality' [11]. More recently, Reinforcement Learning (RL) approaches have been proposed for hybrid diesel engine and battery road vehicles. Hu et al. [12] and Wu et al. [13] implemented Deep Q Network (DQN) to generate EMS for standard driving cycles. It is worth noting that, using a limited number of driving cycles to train an RL agent could lead to the generated EMS only performing satisfactorily under specific driving cycles. Xiong et al. [14] proposed solving the optimal power split problem using Q-learning with Kullback-Leibler divergence, indicating if and when to update the EMS over time. Their results suggest that updating the EMS over time may further reduce fuel consumption. However, such an approach requires frequent updates of the EMS, and the updated EMS may not perform as expected under non-predicted future load cycles.

For shipboard applications, it is rare to find an intelligent EMS which can accommodate non-predicted future voyages. Kalikatzarakis et al. [15] presented "Equivalent Consumption Minimisation Strategies" for shipboard applications with diesel engine in hybridisation with battery and shore power. Their results indicate that a 6% fuel saving can be achieved compared to the rule-based method. Bassam et al. [16] proposed a multi-scheme EMS with a mix of several sub-EMSs at different states for a hybrid fuel cell passenger ship based on an eight-hour profile. Choi et al. [17] implemented a load-following EMS for their hybrid fuel cell and battery boat, in which the fuel cells operate at a designated power setting while the batteries provide any additional power demand. Han et al. [18] proposed a rule-based EMS tuned by a typical load cycle for a low power passenger ferry. It should be noted that the actual EMS performance was not clear for other load cycles since a limited number of power profiles were analysed in these studies.

The research efforts mentioned above have been used to successfully develop EMS for hybrid propulsion systems. However, the existing strategies were tuned with a limited number of load profiles. The actual performance of the strategy for unknown future load profiles, for both road and ship applications, is not clear in the existing works. Although novel approaches such as RL have been applied to develop intelligent EMS, existing EMS are often tuned using a limited number of load profiles. There is a lack of generic EMS that can accommodate stochastic power profiles with high variations in marine applications.

The remainder of this article proposes a novel cost-effective EMS that uses reinforcement learning for shipboard plug-in hybrid fuel cell and battery propulsion systems. The optimal energy management problem will be modelled as a Markov Decision Process (MDP). The formulated MDP is solved through the use of a Double Q reinforcement learning agent utilising real ship stochastic power profiles collected using continuous monitoring. By using continuous monitoring data on a significant scale, the policy generated by the RL agent can be used directly by ships as a guide strategy of the EMS to achieve long-term near-optimal cost performance without prior knowledge of any future power demands.

The rest of this article is organised as follows. Section 2 details the modelling of the hybrid PEMFC and battery propulsion system. Section 3 details the parameters of the candidate ship to which the proposed EMS will be applied. Section 4 formulates the optimal energy management problem using reinforcement learning, and introduces the associated Double Q reinforcement learning agent. Section 5 details the agent training process to acquire the EMS. Section 6 presents the results by applying the Double Q EMS to training and validation voyages. Section 7 details the authors' conclusions.

2. Plug-in hybrid PEMFC and battery propulsion system

2.1. System overview

The work in this article is a continuation of the research detailed in [19]. In that work, the hybrid plug-in PEMFC and battery propulsion system sizing has been optimised based upon a proposed hybrid propulsion system model for a specific vessel operating on a particular route. Figure 1 provides the quasi-steady-state model which has been developed and validated in [19] for system design optimisation and intelligent EMS development. The system comprises a fuel cell, battery and converters. The EMS manages the power flows between the power sources and the load. The battery can discharge or be charged through the bidirectional DC/DC converter.

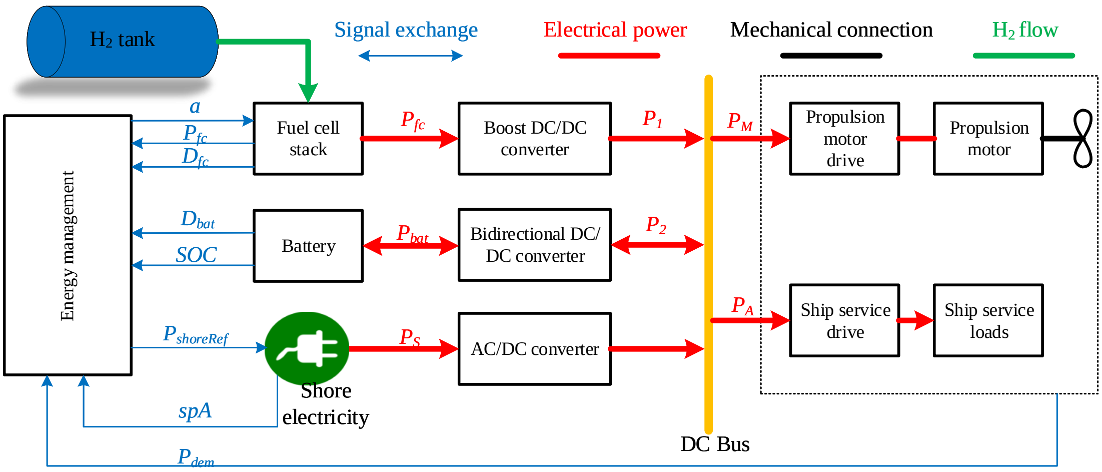


Figure 1: Plug-in hybrid PEMFC and battery propulsion system top-level model. The energy management system manages the power flows within the hybrid system.

2.2. System model

2.2.1. Power converters

The efficiencies of the power converter for the fuel cell (η_1) and the two converting modes of the battery power converter (η_2 for dischraging and η_3 for charging) are presented in Figure 2 [19]. The efficiencies of the power converter models are calculated as the percentage ratios of input to output power in per unit terms. Note that the bidirectional DC/DC converter works at different efficiencies in battery discharging and charging modes.

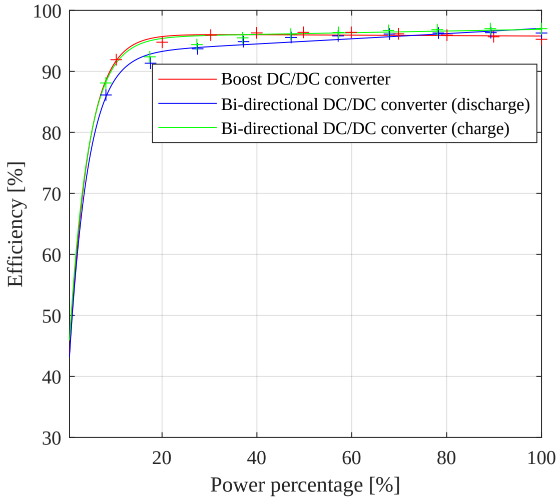


Figure 2: Power converter efficiencies for the PEMFC and battery. Note that the converter efficiency models are representative of achievable characteristics, and can be updated when actual converter features are available for actual engineering applications.

2.2.2. PEMFC model

The input to the fuel cell model is the fuel cell per unit power, denoted by p_{fc}. The specific H_2 consumption is calculated by the model as detailed in Figure 3. The H_2 mass flow rate, $\dot{m}_{\mathrm{H}_2}$, can be calculated by:

$$\dot{m}_{\mathrm{H}_2} = \psi_{fc}(p_{fc})P_{fc} \tag{1}$$

where ψ_{fc} denotes the model function, P_{fc} is the fuel cell rated power. PEMFC degradation is subject to factors such as power transients, load levels and cycling. In this study, the fuel cell degradation characteristics are based upon the work of Chen et al. [20], which have been adopted in [10] for EMS development. At each time step, the cost incurred through fuel cell degradation is $\delta_{fc}P_{fc}\sigma_{fc}$, where δ_{fc} is the inverse of the total available fuel cell operating time steps under current operating conditions, P_{fc} is fuel cell rated power and σ_{fc} is the fuel cell price per kW.

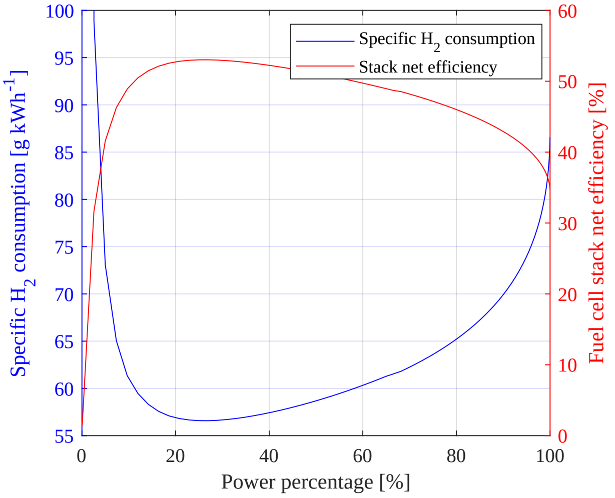


Figure 3: PEMFC specific H_2 consumption and system efficiency. The PEMFC model outputs specific H_2 consumption based on per unit power input.

2.2.3. Battery model

As shown in Figure 4a, the scalable battery model is developed by connecting the outputs from individual cells in series and parallel to form the battery stacks [21]. Individual cell open-circuit voltage is a function of battery State of Charge (SOC) as presented in Figure 4b. The battery model is calibrated and validated using experimental data from [22]. It has been assumed that all the battery cells have identical characteristics and the resistances between cells are negligible. An averaged battery degradation function δ_{bat} is adopted, i.e. the batteries are assured to operate for a fixed period before reaching the end of life. Note that δ_{bat} is the inverse the of total available battery operating time steps.

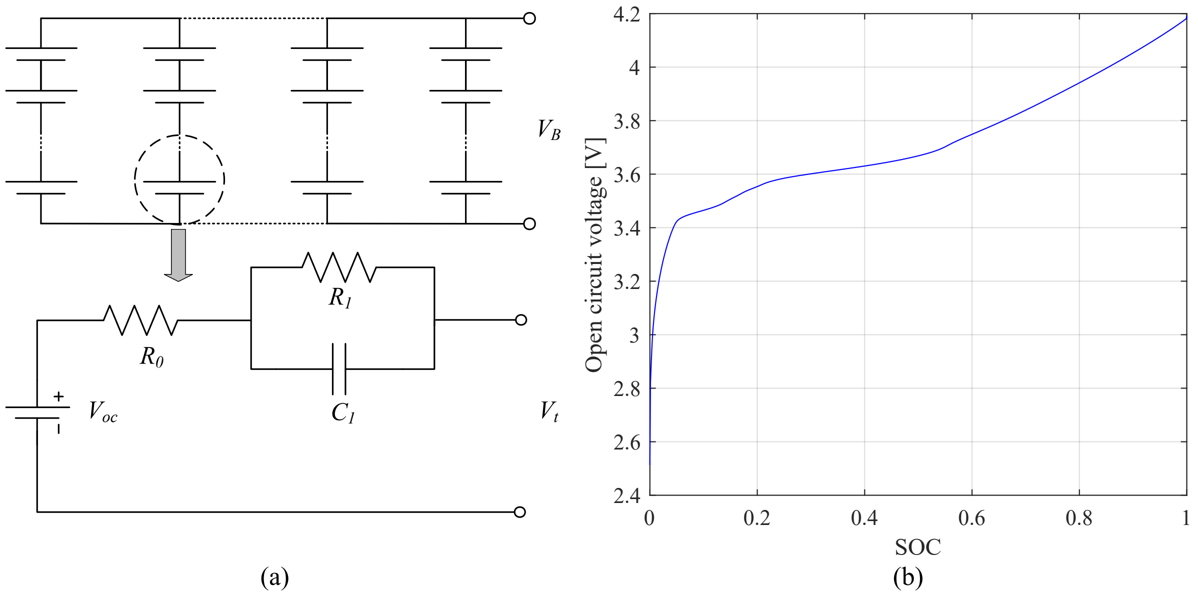


Figure 4: Battery stack model and individual cell circuit (a): the outputs from individual battery cells are connected in series and parallel to form the battery module, and (b) battery individual cell voltage in a function of battery SOC.

The battery works either in discharging/floating or charging modes. The battery output power is:

$$P_{bat} = I_B V_B \tag{2}$$

where I_B and V_B are battery module current and terminal voltage respectively, and V_B is a linear function of individual battery cell terminal output voltage V_t. The cell terminal voltage V_t is calculated by:

$$V_t = V_{oc} - i_0 R_0 - V_c \tag{3}$$

where V_{oc} is a function of battery SOC as shown in Figure 4b. And the voltage across C_1 satisfies:

$$\dot{V}_c = -\frac{V_c}{R_1 C_1} + \frac{i_0}{C_1} \tag{4}$$

The battery SOC change over a period t_1 to t_2 can be calculated by:

$$\Delta SOC = -\eta_b \int_{t_1}^{t_2} C(t)dt \tag{5}$$

where $C(t)$ is battery C-rate, η_b is battery coulombic efficiency.

2.2.4. System dynamics

The action of the battery improves system efficiency and performance by load levelling and peak shaving. For each time step, the required battery power is determined by (see Figure 1):

$$P_1 + P_2 + P_s = P_{dem} \tag{6a}$$

$$P_1 = P_{fc}\eta_1 \tag{6b}$$

$$P_{fc} = \begin{cases} \geq 0, & \text{sailing mode} \\ = 0, & \text{port mode} \end{cases} \tag{6c}$$

$$P_s = \begin{cases} = 0, & \text{sailing mode} \\ \geq 0, & \text{port mode} \end{cases} \tag{6d}$$

$$P_2 = \begin{cases} P_{bat}\eta_2, & P_{bat} \geq 0 \text{ for battery discharging or floating} \\ P_{bat}/\eta_3, & P_{bat} < 0 \text{ for battery charging} \end{cases} \tag{6e}$$

where P_1 is fuel cell power delivered by the boost converter and P_{bat} is battery power, η_1, η_2 and η_3 are the fuel cell and battery converter (charging and discharging) efficiencies respectively, P_{dem} is the total power demand, P_s is the shore power which is only applicable when the vessel is in port.

2.3. Energy management system

The EMS determines the fuel cell power output change through measured inputs of the fuel cell, battery and shore power states as well as power demand (Figure 1). The proposed alternative hybrid propulsion system operates in two modes, i.e. sailing and port modes:

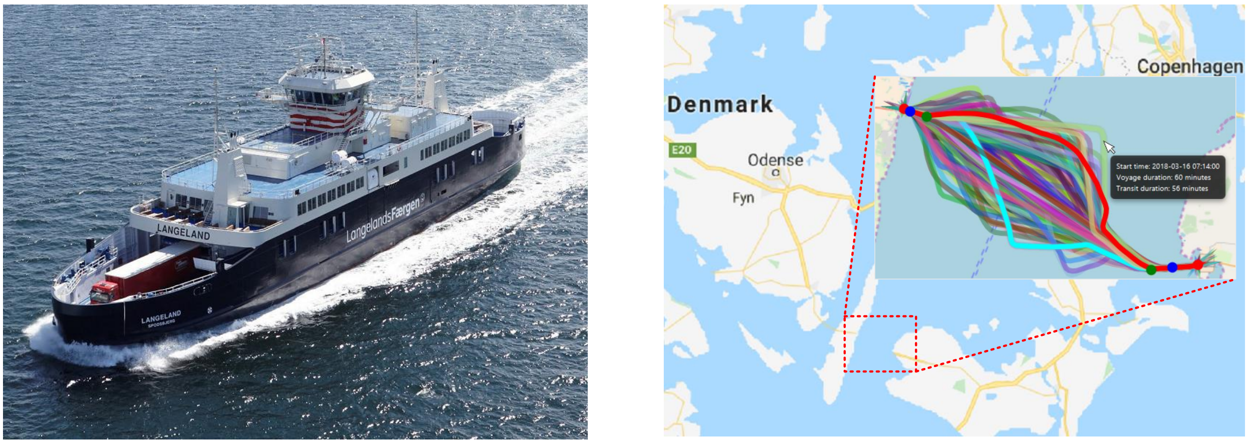
- in port mode, the EMS sets the fuel cell power output to zero, while the shore connection powers the ship's load and recharges the battery;

- in sailing mode, the shore power connection is no longer available–the fuel cell and battery work together to power both propulsion and auxiliary loads. The battery can either work in charging or discharging modes.

3. The candidate ship and continuous monitoring data

The candidate ship and its route are shown in Figure 5. The historical power profiles applied for the agent training were acquired from [5] (1081 voyages in total, from 1 July 2018 to 31 August 2018). Another dataset (391 voyages in total) collected over a different period (from 1 September 2018 to 30 September 2018) will be used for EMS validation. The datasets were first segregated into voyages determined by the ship's speed and location. It should be noted that the power profiles include both propulsive and auxiliary loads.

The length of the original time step of the power profiles is 15 s and remains unchanged. The original power values were smoothed with a Gaussian-weighted moving average filter to reduce measurement noise. The moving average window of the Gaussian filter is 4, and the standard deviation is calculated from 1/5 of the total window width. Due to the large time step and the main focus of this study being energy efficiency and emissions—instead of dynamic performance, the direct current internal resistance ($R_0 + R_1$) of the battery model is used in the subsequent simulations [23].


LANGELAND
LangelandsFærgen
LANGELAND
Copenhagen
Denmark
E20
Odense
Fyn
Start time: 2018-03-16 07:14:00
Voyage duration: 60 minutes
Transit duration: 56 minutes

(a) (b)

Figure 5: Candidate ship (a) and its route (b). The candidate ship is a passenger ferry with integrated full electric propulsion system, and operates between two fixed ports.

On the candidate ship the original diesel-electric system will be replaced by a plug-in hybrid PEMFC and battery system, as described in Section 2. The ship's specifications are presented in Table 1. The original system featured an integrated full electric propulsion configuration with a total installed diesel engine power capacity of 4370 kW. The ship operates between two fixed ports with 8 round trips (16 voyages) per day—each voyage takes approximately 1 h [5].

Table 1: Candidate ship specification.

Parameters	Value
Ship type	Ferry
Gross tonnage	4500
Power system configuration	Integrated full electric propulsion
Installed engine power	4370 kW
Fuel tank volume	140 m^3
Round of trips	8
Average voyage time	1 h

For this research the ship's battery can be recharged in both ports, and the shipboard H_2 storage needs to be replenished once per day. The intended fuel cell power and battery capacity for the alternative plug-in hybrid PEMFC and battery propulsion system are 2940 kW and 581 kWh respectively. The system is capable of delivering a regular service power of 4683 kW and a peak power of 6720 kW, corresponding to battery C-rates of 3 and 6 respectively. Note that the system sizing has been optimised in the authors' earlier work [19]. The adopted H_2 Global Warming Potential (GWP), electricity GWP, H_2 price and electricity price are set at 1.5 kg CO_2kg^{-1}, 0.166 kg CO_2kWh^{-1}, 8.240 $\$kg^{-1}$, and 0.089 $\$kWh^{-1}$ respectively. The battery SOC upper and lower limits are limited to upper and lower values of 0.90 and 0.25 respectively, and the maximum C-rate is 6. Note that the SOC limits are soft constraints, meaning they can be exceeded if this is deemed as necessary. The battery needs to be charged to a SOC of 0.9 prior to departure. A starting SOC of 0.90 affords the system the flexibility to excessive power from the fuel cells if and when required. SOC below 0.25 should be avoided to provide minimum charge reservation, as well as to extend battery life [24].

4. Reinforcement learning based energy management system

This section formulates the optimal energy management problem of the plug-in hybrid PEMFC and battery system with MDP and introduces the Double Q RL agent which will be applied to solve the formulated MDP. Table 2 summaries the RL terminologies, which will be used for the optimal energy management problem in the subsequent sections. Note that a policy learned by a RL agent will be called as the energy management strategy of an energy management system.

Table 2: Summary of RL terminologies in the optimal energy management problem.

Terminology	Description
Agent	Double Q algorithm
Environment	Hybrid system model and historical power profiles
States	System states, including current PEMFC power level, battery SOC, power demand and shore power availability
Action	Fuel cell power change
Reward	A function of constraints and cost incurred in one time step
Policy	Energy management strategy of EMS

Figure 6 shows the detailed MDP agent-environment interaction framework for energy management problem. The environment of the MDP framework includes the hybrid PEMFC and battery system model and historical voyage data [5].

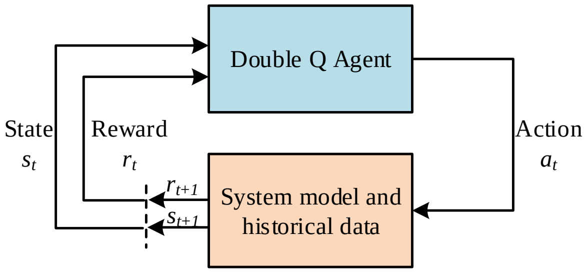


Figure 6: MDP agent-environment interaction framework. The environment of the framework consists of the hybrid PEMFC and battery system model and historical load profiles collected via continuous monitoring.

4.1. Markov Decision Process

An MDP is a stochastic control process in discrete time space, which provides a mathematical framework to model sequential-decision making problems. Such a process can be represented by a tuple (S, A, P, R), where S is a finite set of states, A is a finite set of actions, P is the state transition probability, i.e. $P_{ss',a} = \mathbb{P}[S_{t+1} = s^{'}|S_t = s, A_t = a]$, and r is a reward function $r_{ss',a} = \mathbb{E}\left[r_{t+1}|S_t = s, A_t = a\right]$. The action-value function, which is also called the Q function, for an episodic task with finite horizon of T, is the expected return of taking action a in state s following a policy π:

$$Q(s, a) = \mathbb{E}\left[\sum_{k=0}^{T} \gamma^k r_{t+k}|s_t = s, a_t = a\right] \tag{7}$$

Solving an MDP is to find a optimal policy π^*:

$$\pi^*(s) = \arg\max_a \mathbb{E}\left[\sum_{k=0}^{T} \gamma^k r_{t+k}|s_t = s, a_t = a\right] \tag{8}$$

which leads to the optimal action-value function [11]:

$$Q^*(s, a) = \max_\pi \mathbb{E}\left[\sum_{k=0}^{T} \gamma^k r_{t+k}|s_t = s, a_t = a\right] \tag{9}$$

where $\gamma \in [0, 1]$ is the discount rate. As shown in Figure 6, at time step t, in current state s_t, the agent takes action a_t under the policy π, and observes the resulting next state s_{t+1} and immediate reward r_{t+1} returned from the environment.

4.1.1. States

In the optimal energy management problem, the states represent the current system status. In the proposed system, such states are characterised by shore power availability, spA, system power demand, P_{dem}, fuel cell power level $x \in [0, 1]$, and battery $SOC \in [0, 1]$. spA is binary, i.e. $spA = 0$ when the ship is sailing and $spA = 1$ when the ship is in port. It is assumed that the transition from transit to port can be done instantly, i.e., the shore power is immediately available when the ship is in port. Although x, SOC and P_{dem} are continuous physical parameters they are however divided into discrete grids. Such that the gridded state space can be formulated by looping through all possible state combinations. As each of the state parameters has a finite dimension, the total number of states is the product of the four state dimensions. Each possible state is assigned with a unique state index sequentially (i.e. from 1 to the total number of states). At time step t, the exact state of the system:

$$s_{actual}(t) = [spA(t), P_{dem}(t), x(t), SOC(t)]^T \tag{10}$$

is converted into state index $s(t)$, which is an integer. Note that the environment knows the actual state $s_{actual}(t)$ and $s_{actual}(t+1)$ which results from taking action $a(t)$, but only communicates with the agent using state indices. Such a communication format is designed intentionally so that the agent can record the learning process into tables.

4.2. Action space

In reinforcement learning, the agent interacts with the environment by taking actions in various systems states. The action taken by the agent is the control of fuel cell power change in one time step in this study. The action space is defined as a tuple of possible fuel cell power level changes:

$$A = [a_1, a_2, ..., a_m, ..., a_{n-1}, a_n]^T \tag{11}$$

where $a_1 < 0$ is the maximum decrease and $a_n > 0$ is the maximum increase of fuel cell power output in a time step, $a_m = 0$ indicates fuel cell output power is unchanged and remains constant; all other values of a represent changes of power within the range of (a_1, a_n). The environment overrides an action when the resulting fuel cell power would be negative or greater than the rated power. When action $a_t \in A$ is chosen from the action space at time step t, the fuel cell power level at $t+1$ will be:

$$x_{t+1} = \begin{cases} 0, & x_t + a_t < 0 \\ 1, & x_t + a_t > 1 \\ x_t + a_t, & \text{else} \end{cases} \tag{12}$$

4.3. Reward

The environment returns reward signal r_{t+1} to the agent when action a_t is taken by the agent. The value of r_{t+1} represents how cost-effective a_t is at state s_t:

$$r_{t+1} = \begin{cases} -1, & \text{if } s_{t+1} \text{ is infeasible} \\ -1, & p_{fc} + a_t \notin [0,1] \\ \tanh\left(\frac{1}{cost_{t+1}}\right), & \text{else} \end{cases} \tag{13}$$

where the negative reward of -1 means the agent is penalised if the next state is not feasible or fuel cell power override will occur; the $\tanh\left(\frac{1}{cost_{t+1}}\right)$ function normalises the cost $cost_{t+1}$ to a reward signal in the range of $[0, 1]$ elsewhere. Note that the next state is not feasible if the battery is over charged/discharged or C-rate exceeds the system limit or fuel cell power is not reduced to zero when the ship is in port (fuel cells are not switched off to avoid unnecessary start/stop cycling degradations). $cost_{t+1}$ is the cost incurred in one time step Δt due to action a_t if the next state is feasible:

$$cost_{t+1} = \psi_{fc}(x_t + \frac{a_t}{2})P_{fc}\Delta t\sigma_{H_2} + \delta_{fc}(x_t + \frac{a_t}{2})P_{fc}\sigma_{fc} + P_{sh}\Delta t\sigma_e + \delta_{bat}B\sigma_{bat} \tag{14}$$

i.e. the sum of H_2 cost, fuel cell degradation cost, battery average degradation cost and shore power cost (only when the ship is in port), where σ denotes price. The sub-scripts H_2, fc, e and bat denote H_2, fuel cell, electricity and battery prices respectively. Note the cost $cost_{t+1}$ is unpenalised since the negative reward -1 includes a penalty. To better understand the impact of nonfeasible actions, a penalised cost is also introduced in the following case study. The penalised cost is $cost_{t+1} + 1$ whenever the next state is not feasible, and agent action is overridden or early termination occurs.

4.4. Double Q-learning agent

Q-learning, proposed by Watkins [25], is a model-free approach for solving MDPs, i.e. transition probabilities P are not considered directly during agent training. It is also an off-policy RL method, i.e. the action-values are updated using the next state and the greedy action. When updating the action-value function, the agent acts greedily by choosing an action maximising the next action-value function:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma Q(s\text{'}, \arg\max_{a} Q(s\text{'},a)) - Q(s,a) \right] \tag{15}$$

where s' is the next state. However, the maximisation operations involved in the construction of policy and the ε-greedy action selection processes can lead to poor learning performance as a result of maximisation bias in stochastic environments [26].

This study takes advantage of Double Q-learning (a variant of Q-learning) to learn optimal energy management policies for the sequential power split problem between multiple power sources [11]. Algorithm 2 in Appendix A shows the Double Q-learning agent [27]. The Double Q agent reduces the maximisation bias by using two action-value estimates, Q_1 and Q_2. For each update, with 0.5 probability, Q_2 is used to determine maximising action while Q_1 updates its value:

$$Q_1(s,a) \leftarrow Q_1(s,a) + \alpha \left[r + \gamma Q_2(s\text{'}, \arg\max_{a} Q_1(s\text{'},a)) - Q_1(s,a) \right] \tag{16}$$

Otherwise Q_2 is updated with Q_1 and Q_2 switched. Both the learning rate α and ε of the ε-greedy policy decrease linearly with the increase of learning episodes, and stabilise at fixed values after rate decaying episode number N_d.

4.5. Environment

The environment of the reinforcement learning comprises two parts, i.e. the hybrid propulsion system model [19] and historical power profiles collected using continuous monitoring of operational power demands [5]. Algorithm 1 depicts how the environment of the optimal energy management problem is formulated. Using the historical voyage power profiles, in each learning episode, the environment randomly samples one power profile from the historical data with which the agent interacts. Note that the environment would carry out early termination of an episode if the agent fully discharges the battery ($SOC < 0$) or over-charges the battery ($SOC > 1$). Normal termination occurs when the final targeted time step has been reached. An episode is successful if the agent manages to achieve all the required time steps and recharge the battery to a SOC of soc_H to be fully prepared for next voyage; otherwise, the episode terminates and is recorded as having failed.

Algorithm 1 Environment of the optimal energy management problem

```
Store historical voyage power profiles
for each learning episode do
    Randomly select one sample voyage from historical voyages
    Initialise initial state parameters: $p_{fc} = 0$, $SOC = SOC_H$, $spA = 0$
    for $t = 1 : T$ do
        With action input $a_t$ from the agent, at state $S$ indexed as $s_t$
        Update the next state parameters and the next state index $s_{t+1}$
        Calculate the immediate reward $r_{t+1}$
        if $s_{t+1}$ is infeasible or override happens then
            $r_{t+1} \leftarrow -1$
        else
            $r_{t+1} \leftarrow \tanh\left(\frac{1}{cost}\right)$
            if $t + 1$ is final time step and $SOC_{t+1} = SOC_H$ then
                $r_{t+1} \leftarrow r_{t+1} + 1$
            end if
        end if
        Determine $Termination$
        if $s_{t+1}$ is infeasible or next time step is final time step then
            $Termination \leftarrow True$
            break
        end if
    end for
end for
```

4.6. Agent training

The objective of the on-line EMS is to minimise the overall voyage cost in an environment that is not pre-known. Such an on-line EMS is intended to manage the power flows within the hybrid power system effectively for future unknown voyages. The learning process is an episodic task. In each episode, the environment randomly samples one of the historical voyage power profiles for the agent to interact with to learn a policy minimising the voyage cost. This process repeats until the average episode reward converges. Historical power profiles need to be collected before the beginning of agent training procedure. These profiles will be an inherent part of the RL environment. Note that each profile is unique although there are similarities.

The RL agent training and policy application follow the procedure presented in Figure 7. Note that the RL training parameters such as the learning rate α and the probability of exploration ε at a time step require careful tuning to achieve a strategy with adequate performance:

- the agent should be able to complete the training voyages without early terminations.
- achieve minimum voyage cost with minimum constraint violations.

Once the training is converged, the learned policy, i.e. the strategy of the EMS, needs to be validated using a different set of power profiles. In the application phase, a battery over-discharge protection function ensures the battery modules are not over-discharged. This protection mechanism is beyond the MDP agent-environment interaction framework (Figure 6) and is not functional during agent training (see Figure 7). Such that the agent can learn from penalties during training without external interventions. Actions leading to penalties would be avoided due to their lower Q values in corresponding states.

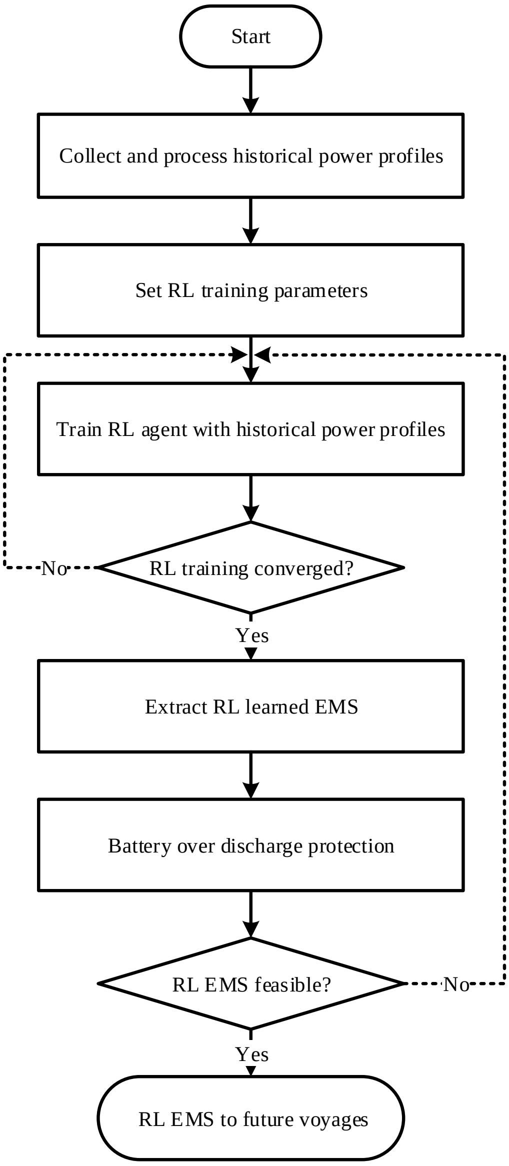

Start
Collect and process historical power profiles
Set RL training parameters
Train RL agent with historical power profiles
RL training converged?
No
Yes
Extract RL learned EMS
Battery over discharge protection
RL EMS feasible?
No
Yes
RL EMS to future voyages

Figure 7: Reinforcement learning agent training and policy application procedure. The RL EMS is acquired with historical power profiles. The battery over discharge protection is enabled in the application phase but disabled during training.

5. Agent training settings

Table 3 shows the grids of state and action spaces for both the Deterministic Dynamic Programming (DDP) strategy (resolution 1) and Double Q strategy. Detailed agent training processes are presented in Appendix B. The DDP implementation is based upon the work of

[28]. The results obtained via DDP are used to evaluate the quality of the strategy generated by the Double Q agent. Therefore, the grids are defined the same in the two algorithms to allow a fair comparison between on-line and off-line strategy initially. Note that developing a strategy by DPP requires complete knowledge of the profile, which is not possible for actual applications. Therefore DDP strategy is only valid as an off-line benchmark to assess the performance of other on-line strategies.

Table 3: State and action space grids.

Parameter	Grid length	Range	Unit
Power demand	50	0–4400	kW
SOC	0.05	0–1	
Fuel cell power level	0.02	0–1	
Shore power availability	–	0 or 1	
Fuel cell power change fraction	0.02	$[-0.04, -0.02, 0, 0.02, 0.04]$	

To further investigate the potential for cost reduction, the DDP strategy SOC grid length was further refined to 0.0125. However, such a refined SOC resolution was not implemented in Double Q strategy due to 'the curse of dimensionality' [11], which would make the problem impossible to solve with the available computational resources.

6. Results

6.1. *Overview of results*

Table 4 details the two different datasets which will be used in this section. Dataset A is used to train the agent to generate the strategy of the EMS. Once the training of the agent has converged, the strategy is verified by removing the random exploration ε adopted in the training phase. Subsequently, the EMS performance is validated using the dataset B, which have not been applied to the agent in the training phase. The strategy is a 4-dimensional action map over the four state parameters. With the system state observed, the optimal action of fuel cell power control can then be found from the action map.

Table 4: Datasets of load profiles and their purposes. Dataset A is used to train the agent to generate the strategy of the EMS. The EMS is then applied to load profiles in dataset B to validate the EMS performance in unseen voyages.

Dataset	Start date	End date	Voyage number	Purpose
A	01/07/2018	31/08/2018	1081	Training/verification
B	01/09/2018	30/09/2018	381	Validation

To verify the strategy performance learned by the Double Q learning agent, the Double Q strategy was applied directly (without any exploration) to the training voyages with over-discharge protection enabled. Such a process will be referred to as verification in the following content. Applying the strategy to a set of validation voyages will be referred to as EMS validation. Table 5 provides a summary of the sample voyages with low, moderate and high power demands, which will be discussed in the following analysis.

Table 5: Summary of sample voyages.

Category	Profile	Average power	Peak power	Voyage time
		[kW]	[kW]	[s]
Training	Training sample 1	904.2	1615.3	3585
	Training sample 2	1086.3	1836.8	3735
	Training sample 3	2040.3	3320.4	3165
Validation	Validation sample 1	1036.8	1487.0	3555
	Validation sample 2	1167.0	2060.0	3555
	Validation sample 3	1597.8	2752.7	3555

As depicted in Table 6, for the training voyages, the Double Q strategy achieved 96.6% cost minimisation performance of the off-line strategy solved by DDP (knowing complete profiles before solving), both with the SOC grid resolution of 0.05. Note that state space resolution also limits the accuracy of DDP [29]. A refined SOC grid resolution of 0.0125 yields an average voyage cost of $740.0 for the training dataset. The Double Q strategy achieves 89.0% cost minimisation performance of the refined DDP solution. For the validation voyages, similar performance was achieved. The DDP strategy results presented in following strategy analysis sections are all solved with SOC resolution of 0.0125.

Table 6: Double Q and DDP strategy average voyage cost comparison.

	DDP_1 [\$]	DDP_2 [\$]	RL [\$]	DDP_1/RL [%]	DDP_2/RL [%]
SOC resolution	0.0125	0.0500	0.0500	-	-
Training voyages	740.0	803.1	831.8	89.0	96.6
Validation voyages	724.9	789.4	815.0	88.9	96.9

Figure 8 presents the voyage cost achieved by the Double Q strategy in comparison with that solved via DDP, for the training (Figure 8a) and validation (Figure 8b) voyages. The Double Q strategy has achieved satisfactory cost performance (only 3.2% higher than DDP strategy) in validation voyages without prior knowledge of future power demands. Note that some voyages in the training dataset have much higher power demands, yielding a maximum Double Q strategy voyage cost close to \$1600.0.

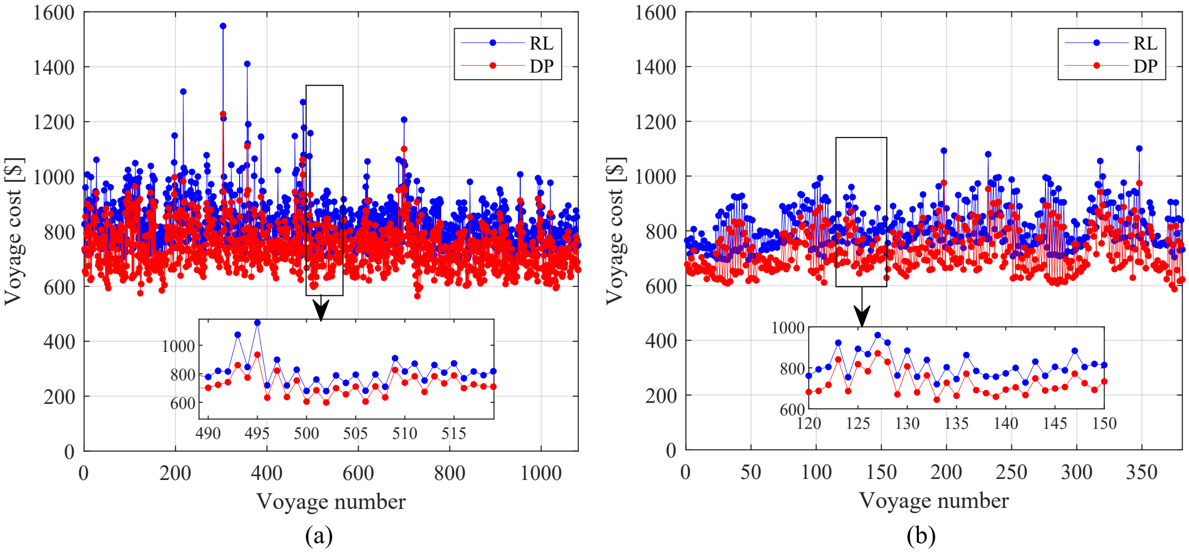


Figure 8: Voyage costs: (a) training voyages and (b) validation voyages. The DDP costs are obtained with a SOC resolution of 0.0125, while it is 0.05 for the Double Q strategy.

6.2. EMS verification

In this section, the Double Q agent generated EMS is applied to three sample voyages in the training dataset. The three sample voyages are with low, moderate and heavy power demands, respectively. Details of the voyage cost and emission compositions are presented. Note that the objective of the EMS is to minimise voyage costs. The voyage emissions are calculated via electricity usage and H_2 consumption with the models presented in [19].

6.2.1. Training sample 1

Figure 9 shows the DDP and Double Q strategies for sample verification voyage 1. This voyage has comparatively lower overall power demands in the training dataset. It starts with relatively higher power demands (1600 kW). During cruising, the power demand stays around 1000 kW. Note that to solve for the DDP strategy requires complete knowledge of the power profiles in advance. The Double Q strategy only takes actions in each time step by observing current system states. The PEMFC power trajectory in the DDP strategy (Figure 9a) is relatively smoother than that of the Double Q strategy (Figure 9b). The Double Q strategy tends to adjust the PEMFC power more frequently within a narrow region, which could be due to the limited knowledge of future power demands. Such behaviour leads to higher PEMFC degradation and H_2 costs (see Table 7). Also, the Double Q strategy rapidly discharges the battery SOC to 0.4 (at 950 s) after departure and then gradually recharges the battery. In contrast, the minimum battery SOC in the DDP strategy is 0.3 and occurs just before shore charging commences (2800 s).

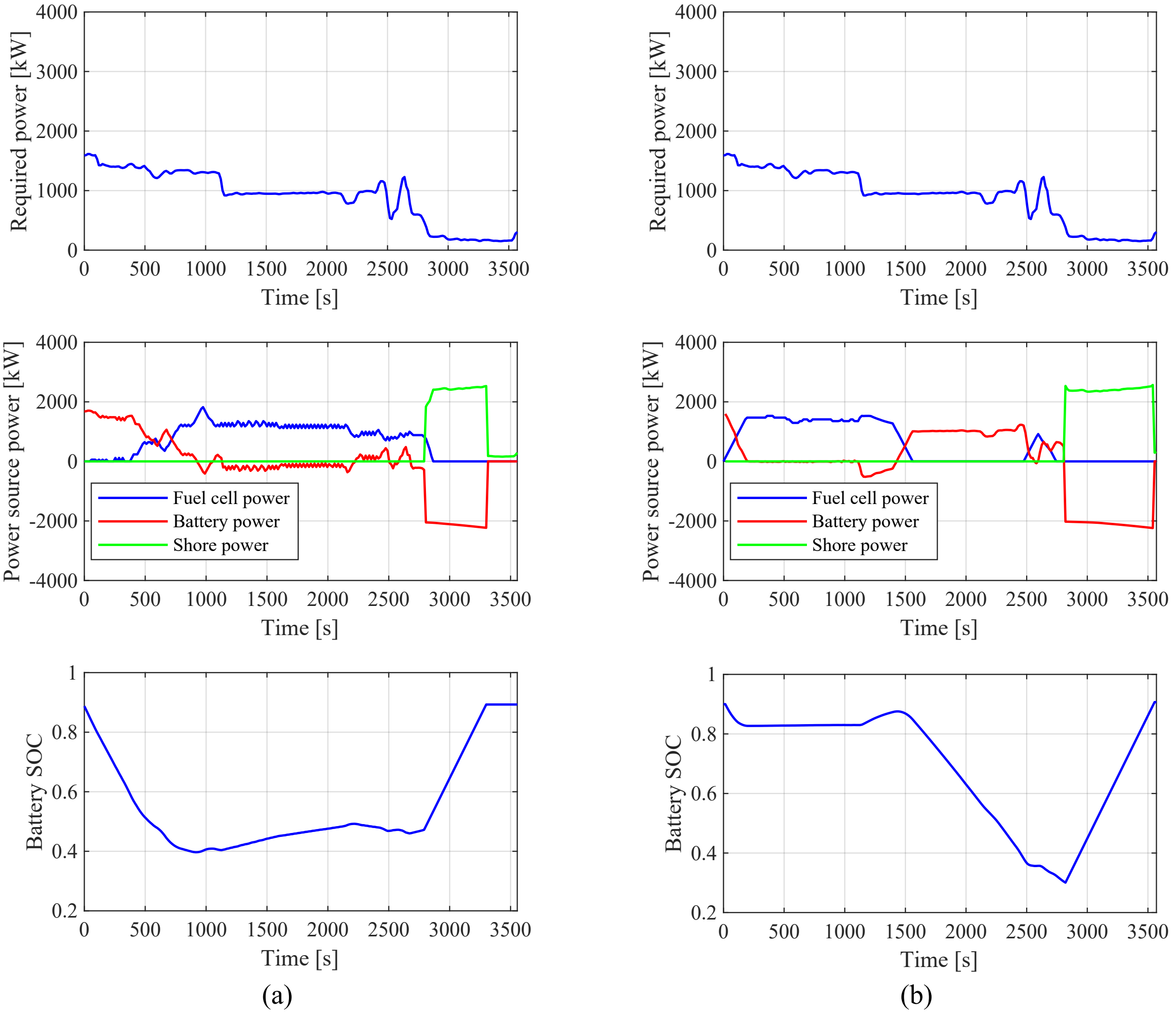


Figure 9: DDP and Double Q energy management strategies for sample training voyage 1 with low power demands: (a) optimal off-line strategy solved by DDP, (b) on-line strategy solved by the Double Q agent.

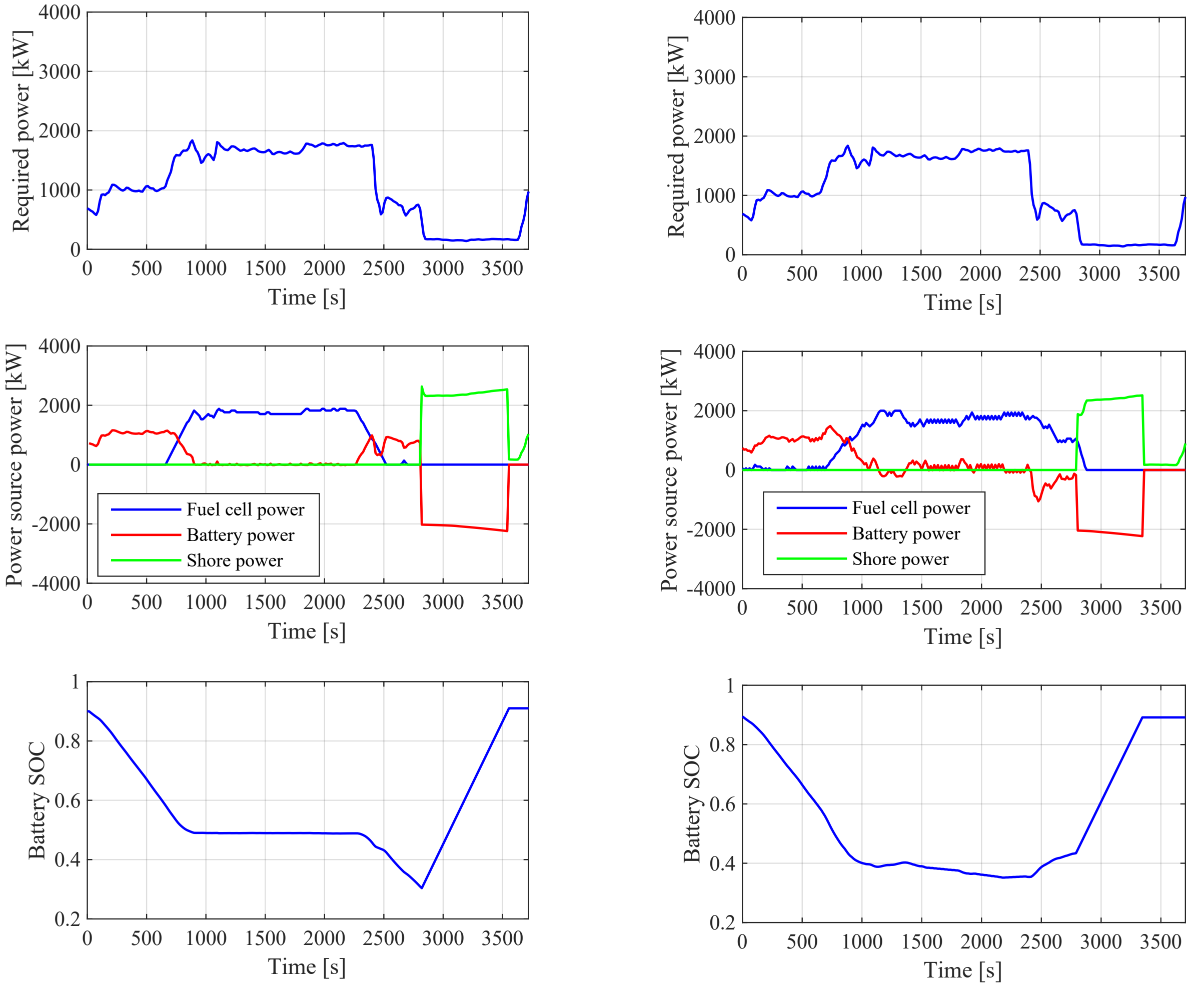
Table 7 details the voyage cost and emission breakdowns of the sample verification voyage 1. The DDP strategy yields voyage cost of \$585.2, which is 85.3% of the Double Q strategy voyage cost. The Double Q strategy leads to higher costs from PEMFC degradation and H_2 consumption. It is worth noting that the voyage GWP emission of the DDP strategy is 11.9% higher than that of the Double Q strategy which is due to the trade-off between voyage cost and GWP emission [19].

Table 7: Double Q and DDP strategy voyage cost and GWP emission breakdown of training sample voyage 1.

	Voyage cost			Voyage GWP Emission		
Training voyage 1	DDP	RL	DDP/RL	DDP	RL	DDP/RL
	[$]	[$]	[%]	[kg]	[kg]	[%]
PEMFC	196.8	252.6	77.9	-	-	-
Battery	64.3	64.3	100.0	-	-	-
Electricity	44.6	31.3	142.4	83.4	58.6	142.4
H_2	279.5	337.7	82.8	50.9	61.5	82.8
Total	585.2	686.0	85.3	134.3	120.1	111.9

6.2.2. Training sample 2

Sample voyage 2 is a typical voyage with moderate power demands in the training dataset. Figure 10 compares the off-line DDP strategy (Figure 10a) and on-line Double Q strategy (Figure 10b) for this voyage. For both RL and DDP strategies, in the departure phase (0-800 s), the batteries provide most of the power from the beginning, while the fuel cells come on line after a delay. The minimum SOC of the DDP strategy for this voyage is approximately 0.25 (at 2850 s). As the RL agent does not exactly know the future power demands and the strategy is generic, the Double Q strategy tends to adjust fuel cell power more frequently. Also, the fuel cells delay being switched to idle until shore power is available, which is because the agent does not know in advance if shore power is available, and the environment was designed to force the fuel cell power to decrease to zero only after shore power was being delivered. Note that, because the ship only stays in port for a short period between voyages, the batteries need to be charged at high C-rates, which could pose additional requirements on the charging infrastructure.



(a)

(b)

Figure 10: DDP and Double Q energy management strategies for sample training voyage 2 with moderate power demands: (a) optimal off-line strategy solved by DDP, (b) on-line strategy solved by the Double Q agent.

Table 8 depicts the cost and GWP emission breakdowns for sample voyage 2 in the training dataset. The Double Q strategy achieves 89.8% cost performance of that of the DDP strategy. Nevertheless, the Double Q strategy yields better GWP emission performance, which has also be observed in sample voyage 1 (Section 6.2.1). The H_2 costs account for 55.4% and 56.3% of the total voyage costs for the DDP and Double Q strategies, respectively. PEMFC degradation costs are the second highest cost source in both strategy results.

Table 8: Double Q and DDP strategy voyage cost and GWP emission breakdown of training sample voyage 2.

	Voyage cost			Voyage GWP Emission		
Training voyage 2	DDP	RL	DDP/RL	DDP	RL	DDP/RL
	[$]	[$]	[%]	[kg]	[kg]	[%]
PEMFC	206.1	246.4	83.6	-	-	-
Battery	67.0	67.0	100.0	-	-	-
Electricity	45.5	34.5	132.1	85.1	64.4	132.1
H_2	395.8	447.5	88.5	72.1	81.5	88.5
Total	714.4	795.3	89.8	157.1	145.9	107.7

6.2.3. Training sample 3

As mentioned in Section 5, the Double Q agent failed to provide a strategy to complete the voyage in less than 0.5% of the training voyages as a consequence of final battery SOC constraint being exceeded. When these failed voyages were examined after the training process it was noted that they had much higher power demands compared to the typical voyages in the training dataset. Figure 11 presents a sample profile when it is known that the ship was heavily laden (corresponds the voyage with maximum cost in Figure 8a), and its optimal EMS solved via DDP (Figure 11a). Unlike the profile discussed in Section 6.2.2, the fuel cell power ramps up immediately after departure for this profile, in contrast to the more normal situation where significant increase in fuel cell power output is delayed as shown in a typical profile similar to Figure 10. Without the battery over-discharge protection, the Double Q strategy tends to discharge the battery rapidly to a SOC below 0.25 after departure from the port. Figure 11b illustrates how the battery over-discharge protection function actuates to minimise the impact and shows how such an override function is effective when tackling voyages with very high power demands.

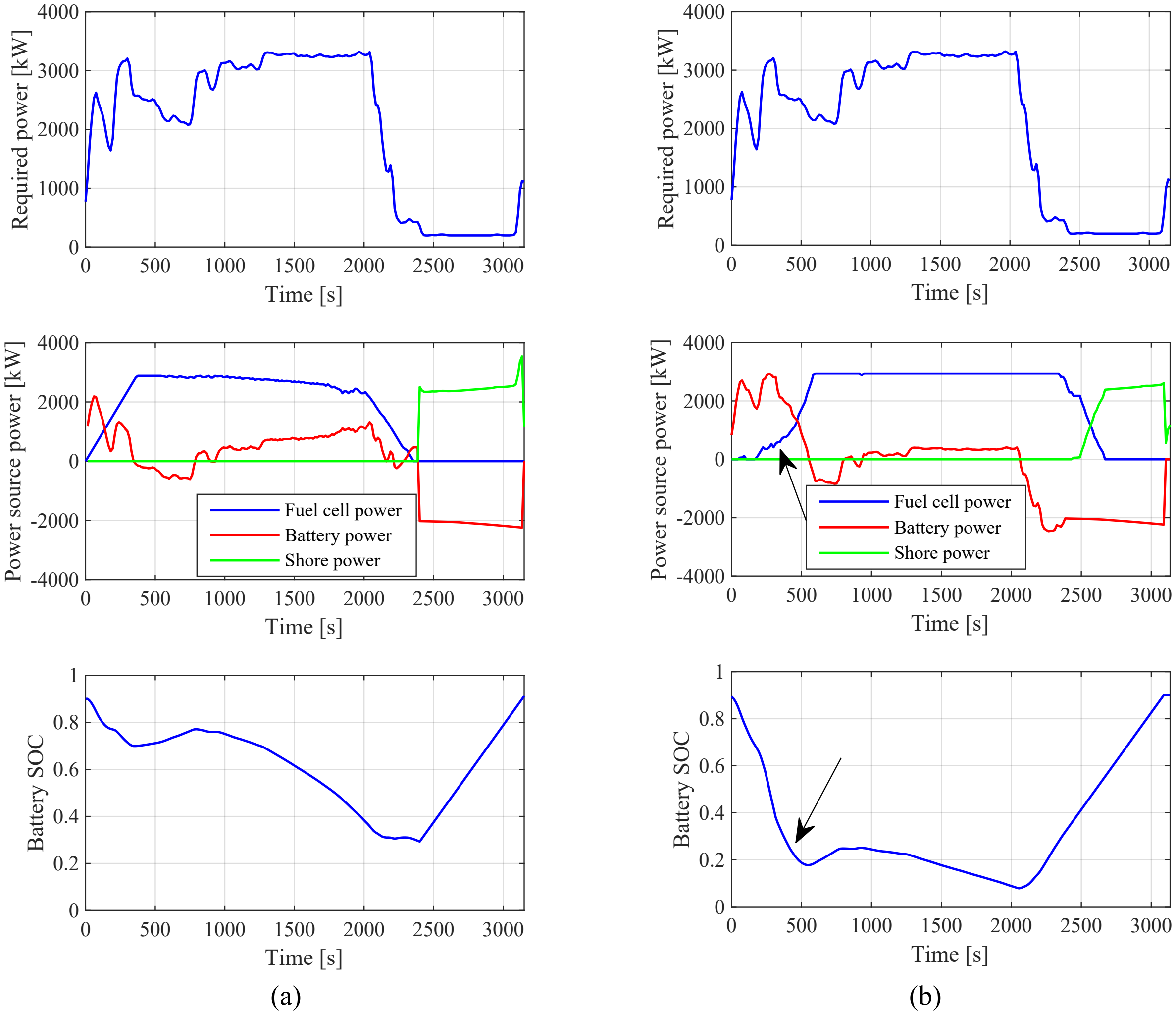


Figure 11: DDP and Double Q energy management strategies for sample training voyage 3 with high power demands: (a) optimal off-line strategy solved by DDP, (b) on-line strategy solved by the Double Q agent.

Table 9 presents a detailed comparison between the DDP and Double Q strategies in terms of voyage cost and GWP emissions. Such a high power profile is unusual in the training dataset. The DDP strategy would generate a voyage cost of \$1228.0, which is 71.8% higher than that of sample voyage 2 (discussed in Section 6.2.2). As a result of the battery over-discharge protection being triggered at 450 s, the PEMFC degradation cost of the Double Q strategy is less than that of the DDP strategy as frequent fuel cell power adjustments have been avoided by action overrides. However, the Double Q strategy outputs a much higher

Table 9: Double Q and DDP strategy voyage cost and GWP emission breakdown of training sample voyage 3.

	Voyage cost			Voyage GWP Emission		
Training voyage 3	DDP	RL	DDP/RL	DDP	RL	DDP/RL
	[\$]	[\$]	[%]	[kg]	[kg]	[%]
PEMFC	242.8	259.0	93.8	-	-	-
Battery	56.7	56.7	100.0	-	-	-
Electricity	46.5	32.9	141.4	87.0	61.5	141.4
H_2	881.9	1199.9	73.5	160.5	218.4	73.5
Total	1228.0	1548.5	79.3	247.5	279.9	88.4

H_2 cost (36% higher), which is due to the PEMFC being forced to run at very high load regions where the fuel efficiency is reduced.

6.3. EMS validation

As the EMS has been developed with the intent to achieve minimum voyage cost for un-predicted future voyages, the trained EMS has been validated by a set of power profiles which have never been included in the training dataset.

6.3.1. Validation sample 1

Figure 12 shows the comparison between the DDP and Double Q strategies of a sample validation voyage with comparatively lower power demands. The Double Q strategy (Figure 12b) discharges the battery modules quickly down to a *SOC* of 0.4 in the first 1000 s, and maintains the fuel cell power output to a narrow region during sailing. The batteries satisfy significant transients in the departing and approaching phases. In contrast, the DDP strategy only discharges the battery rapidly at the beginning of the voyage (0-550 s). Similar trends have been observed in the sample training voyage (Figure 10). The Double Q strategy voyage cost is 12.8% higher than that of the DDP strategy (Table C.1). Nevertheless, the Double Q strategy performs 10.1% better in terms of GWP emission. Such an observation reflects the trade-off between voyage costs and GWP emissions. Note that similar observations have been found in the training sample voyages (see Section 6.2.1 and 6.2.2).

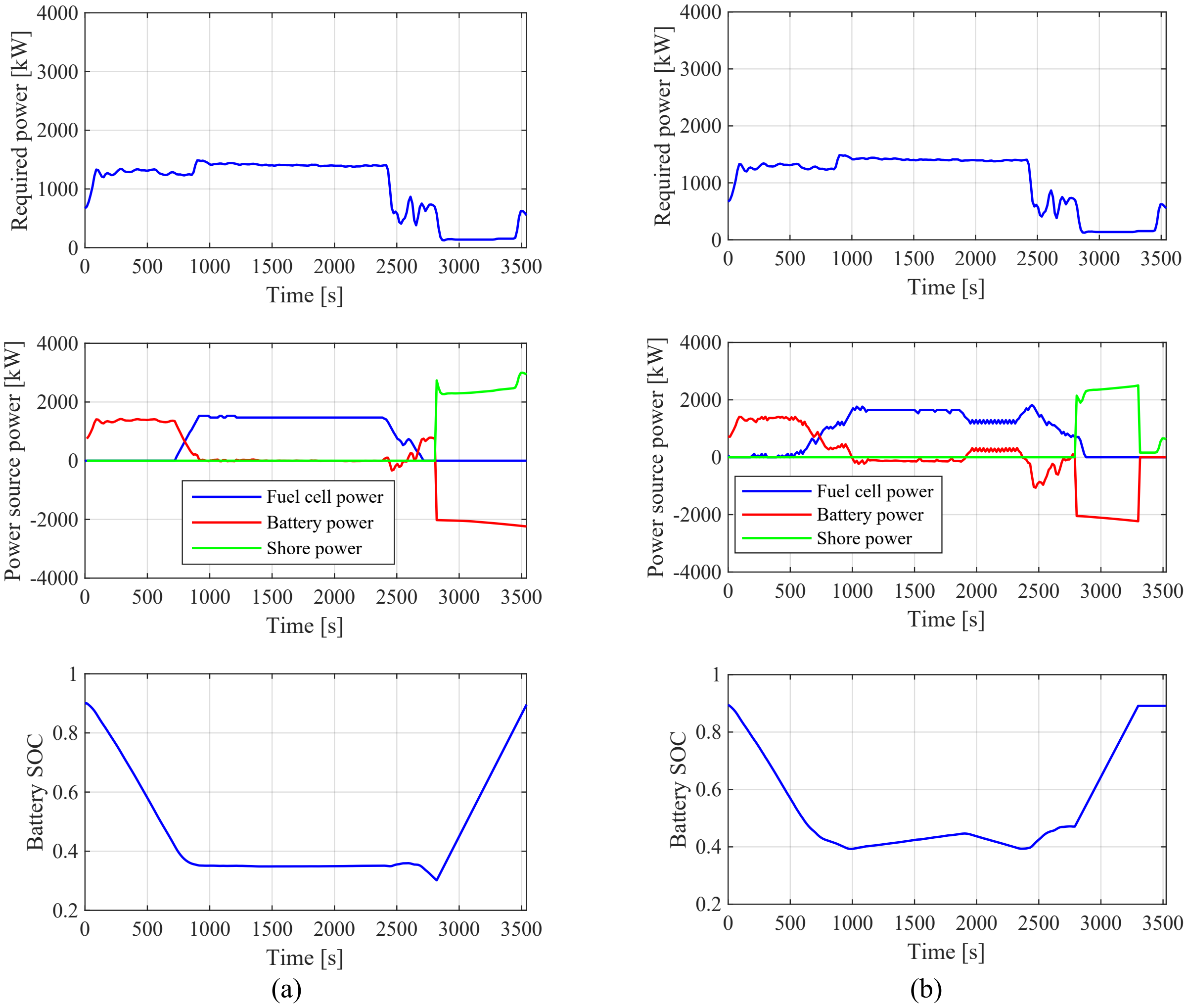


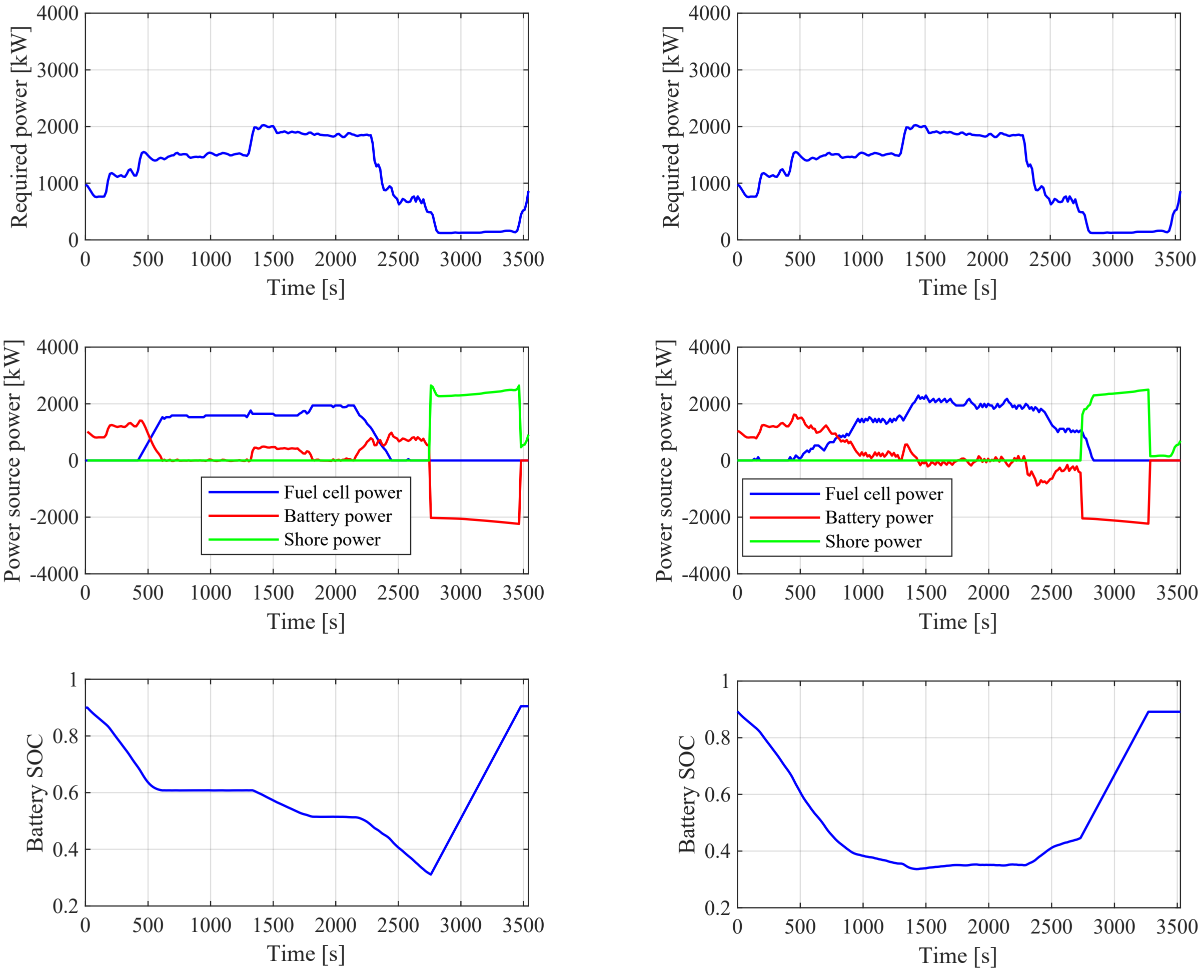
Figure 12: DDP and Double Q energy management strategies for sample validation voyage 1 with low power demands: (a) optimal off-line strategy solved by DDP, (b) on-line strategy solved by the Double Q agent.

6.3.2. Validation sample 2

Figure 13 presents the DDP and Double Q strategies of a sample profile with moderate power demands from the validation dataset. In Figure 13a, as the complete profile is known before solving the DDP strategy, the DDP strategy only adjusts PEMFC power output when necessary. As in Figure 13b, the Double Q strategy adjusts PEMFC power more frequently due to uncertainty regarding the power demands in the next time steps. Such a pattern has also been observed in the first two training sample profiles. The Double Q strategy voyage cost is 11.2% higher than that of the DPP strategy, which is due to frequent PEMFC power adjustments and higher H_2 consumption. Note that the Double Q strategy still performs better than the DDP strategy in terms of GWP emissions (Table C.2).

6.3.3. Validation sample 3

As discussed in Section 6.2.3, the RL agent failed in training voyages with extremely high power demands. Nevertheless, the Double Q strategy managed to complete all the validation voyages without triggering the battery over-discharge protection function. Figure 14 compares the DDP and Double Q strategies. As in Figure 14b, the Double Q strategy discharges the battery rapidly to a *SOC* of 0.4 after departure with a delay before the PEMFC provides any power output. In contrast, the DDP strategy (Figure 14a) ramps the PEMFC output immediately at departure in response to such a high load profile. The voyage cost of the DDP strategy is 89.9% of its RL counterpart (Table C.3). It is worth noting that the GWP emissions produced by the two strategies are very close to each other (0.7% difference). Although the Double Q strategy consumes more H_2 than the DDP strategy, it requires much less shore generated electricity compared to the DDP strategy.



(a)

(b)

Figure 13: DDP and Double Q energy management strategies for sample validation voyage 2 with moderate power demands: (a) optimal off-line strategy solved by DDP, (b) on-line strategy solved by the Double Q agent.

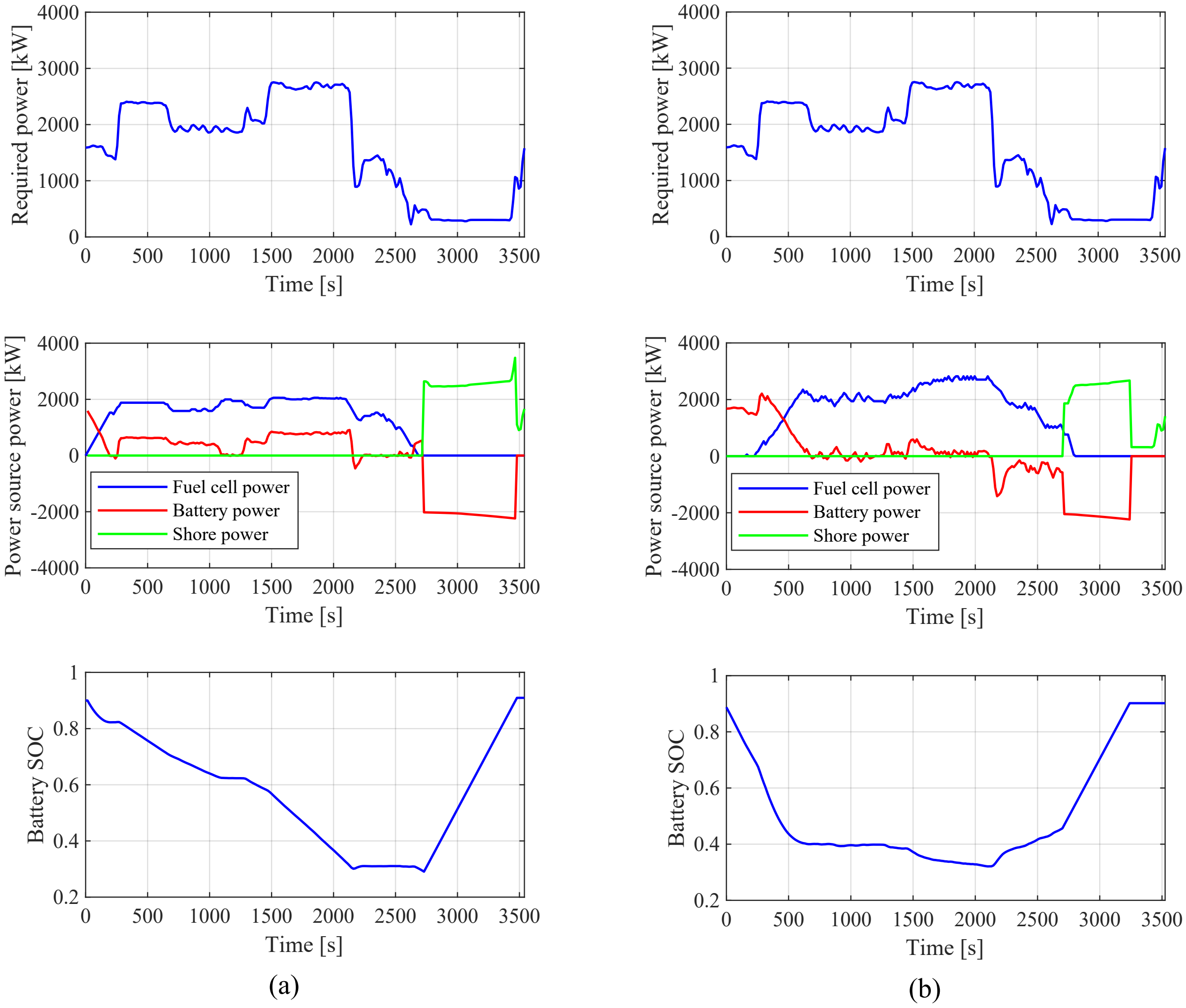


Figure 14: DDP and Double Q energy management strategies for sample validation voyage 3 with high power demands: (a) optimal off-line strategy solved by DDP, (b) on-line strategy solved by the Double Q agent.

6.4. Summary of results

Without prior knowledge of future power demands, the Double Q learning based EMS presented in this article can achieve near-optimal cost performance (96.9%) compared to those solved using DDP with the same state space resolution. Furthermore, the Double Q strategy has achieved average costs which are 12.4% and 12.5% higher than that of the refined DDP strategy across training and validation datasets, respectively. The Double Q agent presented in this article can achieve near-optimal cost performance for the candidate ship in a stochastic environment. Wu et al. [13] reported that, in a non-stochastic environment with a single power profile, their Q learning agent achieved a fuel cost 12.4% higher than that of the dynamic programming strategy in their road vehicle-related study.

Consequently, the proposed EMS can be applied to hybrid fuel cell and battery propulsion system, providing reference signals to the control systems by observing historical and current power demands. Although the objective of the Double Q strategy was designed to minimise voyage costs, due to the trade-off between costs and GWP emissions, the Double Q strategy performs even better than the DDP strategy in terms of GWP emissions (approximately 6% less GWP emissions for both training and validation datasets). More H_2 usage would result in higher voyage costs but lower GWP emissions.

7. Conclusions

This work has formulated and solved the optimal energy management problem of plug-in hybrid fuel cell and battery systems, using the novel approach of Double Q reinforcement learning to achieve near-optimal cost performance to for un-predicted future voyages. Using real ship data collected from the candidate ship via continuous monitoring, the Double Q agent has been trained adequately with a dataset of 1081 training voyages and subsequently validated using another dataset of 381 voyages collected over a separate time period. The approach is novel and can be adopted by hybrid fuel cell and battery ships to achieve near-optimal use of the energy sources. The results show that the Double Q agent can achieve a level of effectiveness similar to that solved by dynamic programming with the identical settings in state and action spaces. Such a similarity indicate that the Double Q agent is effective in dealing with stochastic environments by reducing maximisation biases. Also, such performance suggests that reinforcement learning is a viable approach to solve the optimal power split problem in a hybrid propulsion system, provided that enough historical data has been collected and is made available. In contrast, the Q agent which introduces maximisation biases fails to achieve satisfactory performance, suggesting that the stochasticity of real-world power profiles needs to be properly addressed when developing energy management strategies using reinforcement learning. In future work, the gridded state and action spaces will be extended to continuous spaces with deep neural networks as function approximators to achieve higher resolution.

Acknowledgements

The authors thank Jens Christian Bjeldorf and Molslinje A/S for approving using the ship data in this study. The authors are grateful to Stig Eriksen and his colleagues for collecting the ship data. The authors are also indebted to Konrad Yearwood for his valuable critique of this article.

References

[1] L. van Biert, M. Godjevac, K. Visser, P. V. Aravind, A review of fuel cell systems for maritime applications, Journal of Power Sources 327 (2016) 345–364. doi:10.1016/j.jpowsour.2016.07.007.

[2] D. Larcher, J.-M. Tarascon, Towards greener and more sustainable batteries for electrical energy storage, Nature chemistry 7 (2015) 19. doi:10.1038/nchem.2085.

[3] P. Wu, R. Bucknall, On the design of plug-in hybrid fuel cell and lithium battery propulsion systems for coastal ships, in: P. Kujala, L. Lu (Eds.), 13th International Marine Design Conference (IMDC 2018), volume 2, CRC Press/Balkema, London, 2018, pp. 941–951.

[4] J. J. de-Troya, C. Álvarez, C. Fernández-Garrido, L. Carral, Analysing the possibilities of using fuel cells in ships, International Journal of Hydrogen Energy 41 (2016) 2853–2866. doi:10.1016/j.ijhydene.2015.11.145.

[5] S. Eriksen, M. Lützen, J. B. Jensen, J. C. Sørensen, Improving the energy efficiency of ferries by optimizing the operational practices, in: Proceedings of the Full Scale Ship Performance Conference 2018: The Royal Institution of Naval Architects, The Royal Institution of Naval Architects, 2018, pp. 101–111.

[6] N. Sulaiman, M. Hannan, A. Mohamed, E. Majlan, W. Wan Daud, A review on energy management system for fuel cell hybrid electric vehicle: Issues and challenges, Renewable and Sustainable Energy Reviews 52 (2015) 802–814. doi:10.1016/j.rser.2015.07.132.

[7] Y. Wang, Z. Sun, X. Li, X. Yang, Z. Chen, A comparative study of power allocation strategies used in fuel cell and ultracapacitor hybrid systems, Energy 189 (2019) 116142. doi:10.1016/j.energy.2019.116142.

[8] Y. Wang, X. Li, L. Wang, Z. Sun, Multiple-grained velocity prediction and energy management strategy for hybrid propulsion systems, Journal of Energy Storage 26 (2019) 100950. doi:10.1016/j.est.2019.100950.

[9] P. M. Muñoz, G. Correa, M. E. Gaudiano, D. Fernández, Energy management control design for fuel cell hybrid electric vehicles using neural networks, International Journal of Hydrogen Energy 42 (2017) 28932–28944. doi:10.1016/j.ijhydene.2017.09.169.

[10] T. Fletcher, R. Thring, M. Watkinson, An Energy Management Strategy to concurrently optimise fuel consumption & PEM fuel cell lifetime in a hybrid vehicle, International Journal of Hydrogen Energy 41 (2016) 21503–21515. doi:10.1016/j.ijhydene.2016.08.157.

[11] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, 2018.

[12] Y. Hu, W. Li, K. Xu, T. Zahid, F. Qin, C. Li, Energy Management Strategy for a Hybrid Electric Vehicle Based on Deep Reinforcement Learning, Applied Sciences 8 (2018) 187. doi:10.3390/app8020187.

[13] J. Wu, H. He, J. Peng, Y. Li, Z. Li, Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus, Applied Energy 222 (2018) 799–811. doi:10.1016/j.apenergy.2018.03.104.

[14] R. Xiong, J. Cao, Q. Yu, Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle, Applied Energy 211 (2018) 538–548. doi:10.1016/j.apenergy.2017.11.072.

[15] M. Kalikatzarakis, R. Geertsma, E. Boonen, K. Visser, R. Negenborn, Ship energy management for hybrid propulsion and power supply with shore charging, Control Engineering Practice 76 (2018) 133–154. doi:10.1016/j.conengprac.2018.04.009.

[16] A. M. Bassam, A. B. Phillips, S. R. Turnock, P. A. Wilson, Development of a multi-scheme energy management strategy for a hybrid fuel cell driven passenger ship, International Journal of Hydrogen Energy 42 (2017) 623–635. doi:10.1016/j.ijhydene.2016.08.209.

[17] C. H. Choi, S. Yu, I.-S. Han, B.-K. Kho, D.-G. Kang, H. Y. Lee, M.-S. Seo, J.-W. Kong, G. Kim, J.-W. Ahn, S.-K. Park, D.-W. Jang, J. H. Lee, M. Kim, Development and demonstration of PEM fuel-cell-battery hybrid system for propulsion of tourist boat, International Journal of Hydrogen Energy 41 (2016) 3591–3599. doi:10.1016/j.ijhydene.2015.12.186.

[18] J. Han, J.-F. Charpentier, T. Tang, An Energy Management System of a Fuel Cell/Battery Hybrid Boat, Energies 7 (2014) 2799. doi:10.3390/en7052799.

[19] P. Wu, R. Bucknall, Hybrid fuel cell and battery propulsion system modelling and multi-objective optimisation for a coastal ferry, International Journal of Hydrogen Energy 45 (2020) 3193–3208. doi:10.1016/j.ijhydene.2019.11.152.

[20] H. Chen, P. Pei, M. Song, Lifetime prediction and the economic lifetime of proton exchange membrane fuel cells, Applied Energy 142 (2015) 154–163. doi:10.1016/j.apenergy.2014.12.062.

[21] X. Hu, L. Johannesson, N. Murgovski, B. Egardt, Longevity-conscious dimensioning and power management of the hybrid energy storage system in a fuel cell hybrid electric bus, Applied Energy 137

(2015) 913–924. doi:10.1016/j.apenergy.2014.05.013.

[22] F. Zheng, Y. Xing, J. Jiang, B. Sun, J. Kim, M. Pecht, Influence of different open circuit voltage tests on state of charge online estimation for lithium-ion batteries, Applied Energy 183 (2016) 513–525. doi:10.1016/j.apenergy.2016.09.010.

[23] J. Kim, J. Shin, C. Chun, B. H. Cho, Stable configuration of a li-ion series battery pack based on a screening process for improved voltage/soc balancing, IEEE Transactions on Power Electronics 27 (2012) 411–424. doi:10.1109/TPEL.2011.2158553.

[24] N. Omar, M. A. Monem, Y. Firouz, J. Salminen, J. Smekens, O. Hegazy, H. Gaulous, G. Mulder, P. Van den Bossche, T. Coosemans, et al., Lithium iron phosphate based battery–assessment of the aging parameters and development of cycle life model, Applied Energy 113 (2014) 1575–1585. doi:10.1016/j.apenergy.2013.09.003.

[25] C. J. C. H. Watkins, Learning from delayed rewards, Ph.D. thesis, King's College, Cambridge, 1989.

[26] H. van Hasselt, A. Guez, D. Silver, Deep Reinforcement Learning with Double Q-learning, 2015. arXiv:1509.06461.

[27] H. van Hasselt, Double q-learning, in: J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, A. Culotta (Eds.), Advances in Neural Information Processing Systems 23, Curran Associates, Inc., 2010, pp. 2613–2621.

[28] O. Sundström, D. Ambühl, L. Guzzella, On implementation of dynamic programming for optimal control problems with final state constraints, Oil & Gas Science and Technology–Revue de l'Institut Français du Pétrole 65 (2010) 91–102. doi:10.2516/ogst/2009020.

[29] X. Wang, H. He, F. Sun, J. Zhang, Application study on the dynamic programming algorithm for energy management of plug-in hybrid electric vehicles, Energies 8 (2015) 3225–3244. doi:10.3390/en8043225.

[30] M. Rouholamini, M. Mohammadian, Heuristic-based power management of a grid-connected hybrid energy system combined with hydrogen storage, Renewable energy 96 (2016) 354–365. doi:10.1016/j.renene.2016.04.085.

Appendix A. Double Q RL agent

Algorithm 2 Double Q RL agent [27]

```
$Q_1(s,a)=0$, $Q_2(s,a)=0$, $\forall s \in S$, $\forall a \in A$
$n=1$, $\alpha=1$, $\varepsilon=1$
while $n < N_{max}$ do
    repeat
        if $n \leq N_d$ then
            $\alpha \leftarrow \alpha - \Delta\alpha \times n$
            $\varepsilon \leftarrow \varepsilon - \Delta\varepsilon \times n$
        end if
        if $rand < \varepsilon$ then
            Select action $a$ randomly from $A$
        else
            $a \leftarrow \arg\max_a (Q_1(s,a)+Q_2(s,a))$
        end if
        Take action $a$, observe $r, s'$ and $terminationflag$
        With 0.5 probability updating $Q_1$
        if update $Q_1$ then
            $Q_1(s,a) \leftarrow Q_1(s,a) + \alpha\,[r + \gamma Q_2(s', \arg\max_a Q_1(s',a)) - Q_1(s,a)]$
        else
            $Q_2(s,a) \leftarrow Q_2(s,a) + \alpha\,[r + \gamma Q_1(s', \arg\max_a Q_2(s',a)) - Q_2(s,a)]$
        end if
        $s \leftarrow s'$
    until $terminationflag$ is $true$
end while
```

Appendix B. Agent training process

Table B.1 shows the parameters used to train the Double Q agent. The parameter ε represents the probability of exploration at a time step. The learning rate α determines to what degree the temporal difference is acquired: $\alpha = 1$ suggest that only the most recent information is learned, $\alpha = 0$ nothing new has been learned. Both α and ε decrease linearly from their initial values whilst the training episode number is less than N_s. Such settings reflect the need for the agent to explore less frequently and learn more cautiously when enough experience has been gained, whereas more aggressive and bold learning style is preferred at the beginning to quickly gain experience. As the energy management problem is formulated with an average episode length of 240, and the costs incurred in all steps are of equal importance, the discount rate γ is set at 1 (i.e. un-discounted). It is worth mentioning that careful tuning of these parameters is necessary to balance the conflict between *exploration and exploitation* [11].

Table B.1: Reinforcement learning parameters.

Parameter	Description	Value
α_{init}	Initial learning rate	1.0
$\Delta\alpha$	Learning rate decaying rate	3.3×10^{-6}
ε_{init}	Initial ε	1.0
$\Delta\varepsilon$	ε decaying rate	3.3×10^{-6}
γ	Discounting rate	1.0
N_s	Episode α and ε stabilises	3.0×10^5

Figure B.1 shows the learning process of the Double Q agent. It is interesting that the mean episode reward decreases to -12 initially (0.6×10^5 episodes). This decrease suggests that initially a divergent policy was being learned before the agent was able to learn towards a convergent policy. The training was terminated after 5×10^5 episodes (4.8 h on an Intel i7-4790 processor using single thread in Matlab 2019a).

The mean episode reward stabilised to a value of 88 after about 3×10^5 episodes of training (Figure B.1a), while the maximum episode reward stabilised around 120. Such stabilisation suggests that the algorithm has converged. The average success rates (see Algorithm 1) were close to 100% after convergence. Note that this rate is not exactly 100% (Figure B.1b) which is mainly due to a small exploration probability (1.0×10^{-3}) that still exists and a minor fraction of training voyages with high power demands vary significantly from other voyages. In Figure B.1c, both the actual episode cost and penalised episode cost increase rapidly in the first 1×10^5 episodes. The reason for that is early termination frequently occurs and at the initial stage of the training. In other words, the agent could not complete most training voyages in the initial stages of training (also see the mean episode steps in Figure B.1d) due to the policy's tendency to drain the battery aggressively from the beginning.

As the training goes on, the agent managed to complete most of the training voyages from 2×10^5 episodes onwards. Also, the average voyage cost starts to decrease after 2×10^5 episodes. The actual cost and penalised cost (including the penalties caused by exceeded constraints) overlap with each other, suggesting infeasible actions have been reduced to a minimum. In summary, the agent appears to complete voyages first, then learn to minimise voyages costs (maximum reward) due to the reward setup. In contrast, as shown in Appendix Figure B.2, with the same hyperparameter settings, the Q agent failed to converge to a policy with reasonable performance, owing to the presence of maximisation biases throughout the learning process (see Eq. 15). These biases cause over-estimation of action-value function, which leads to unstable trainings in Q-learning. The Double Q-learning reduces such biases by using two Q-functions.

Note that the environment is highly stochastic, a small fraction of training voyages with high power demands vary significantly from other voyages; the learned policy fails to fulfil the final battery SOC constraint of $SOC = SOC_H$ in less than 0.5% of the 1081 total training voyages. This failure suggests that an override function would be necessary to make the learned policy fully compliant with the final battery state constraint. A battery over-discharge protection, as in Figure 7, was proven to be effective. This protection is realised by forcing the fuel cell power to increase by 5% of rated power in one time step when the battery SOC drops below the lower limit (0.25) [30].

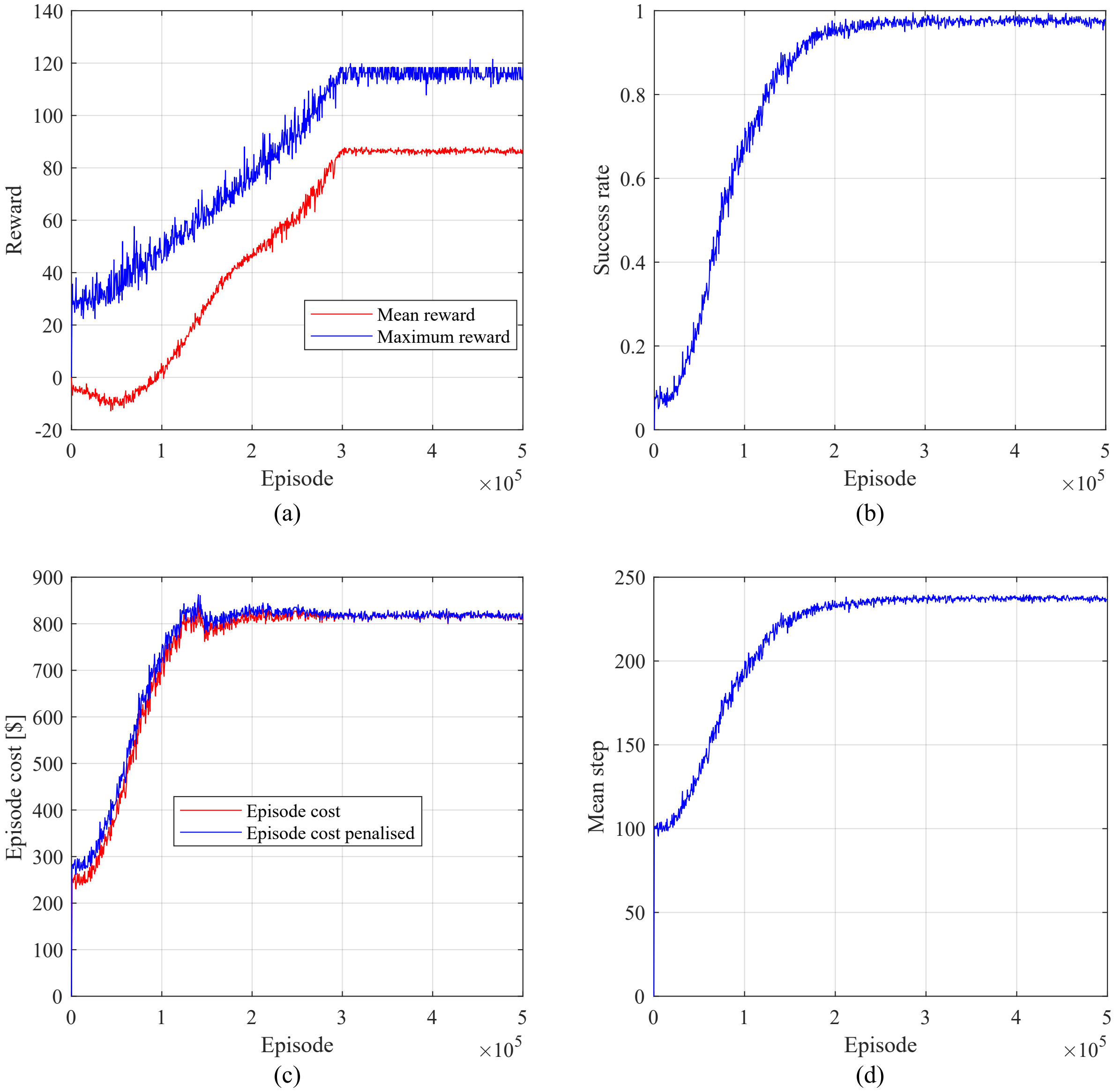


Figure B.1: Double Q agent training process: (a) average reward, (b) maximum reward, (c) average penalised and unspecialised costs and (d) average episode steps of every 500 episodes.

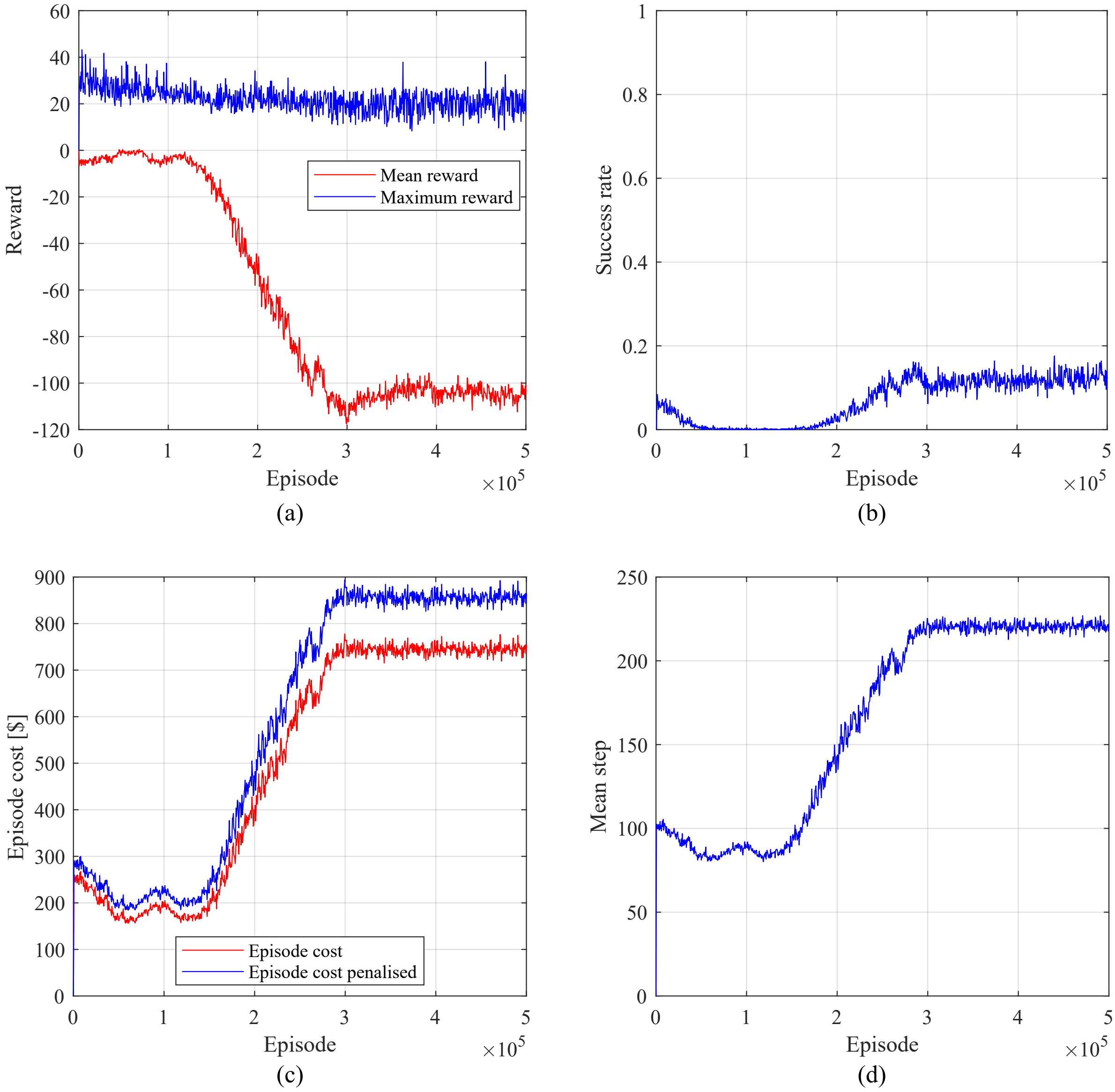


Figure B.2: Q agent training process: (a) average reward, (b) maximum reward, (c) average penalised and unpenalised costs and (d) average episode steps of every 500 episodes. The Q agent failed to converge to a policy with reasonable cost performance and the constraints were violated frequently in late stage of training.

Appendix C. Cost and emission breakdowns of validation sample voyages

Table C.1: Double Q and DDP strategy voyage cost and GWP emission breakdown of validation sample voyage 1.

	Voyage cost			Voyage GWP Emission		
Validation voyage 1	DDP	RL	DDP/RL	DDP	RL	DDP/RL
	[$]	[$]	[%]	[kg]	[kg]	[%]
PEMFC	208.6	244.7	85.2	-	-	-
Battery	63.7	63.7	100.0	-	-	-
Electricity	44.4	31.2	142.2	82.9	58.3	142.2
H_2	345.2	406.8	84.9	62.8	74.1	84.9
Total	661.9	746.5	88.7	145.8	132.4	110.1

Table C.2: Double Q and DDP strategy voyage cost and GWP emission breakdown of validation sample voyage 2.

	Voyage cost			Voyage GWP Emission		
Validation voyage 2	DDP	RL	DDP/RL	DDP	RL	DDP/RL
	[$]	[$]	[%]	[kg]	[kg]	[%]
PEMFC	211.7	239.6	88.4	-	-	-
Battery	63.7	63.7	100.0	-	-	-
Electricity	43.9	32.4	135.6	82.0	60.5	135.6
H_2	411.9	477.6	86.3	75.0	86.9	86.3
Total	731.2	813.2	89.9	157.0	147.4	106.5

Table C.3: Double Q and DDP strategy voyage cost and GWP emission breakdown of validation sample voyage 3.

	Voyage cost			Voyage GWP Emission		
Validation voyage 3	DDP	RL	DDP/RL	DDP	RL	DDP/RL
	[$]	[$]	[%]	[kg]	[kg]	[%]
PEMFC	256.2	257.4	99.5	-	-	-
Battery	63.7	63.7	100.0	-	-	-
Electricity	50.4	36.9	136.3	94.1	69.1	136.3
H_2	605.2	734.9	82.3	110.2	133.8	82.3
Total	975.5	1093.0	89.3	204.3	202.8	100.7