Journal of Power Sources 455 (2020) 227964

Contents lists available at ScienceDirect

Journal of Power Sources

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

Data-driven reinforcement-learning-based hierarchical energy management strategy for fuel cell/battery/ultracapacitor hybrid electric vehicles☆

Haochen Sun [a], Zhumu Fu [a,b,*], Fazhan Tao [a,b,**], Longlong Zhu [a], Pengju Si [a,b]

[a] School of Information Engineering, Henan University of Science and Technology, Luoyang, China
[b] Henan Key Laboratory of Robot and Intelligent Systems, Henan University of Science and Technology, Luoyang, China

HIGHLIGHTS

- Decouple power demand of FCHEV based on fuzzy filter to improve RL algorithm.
- Propose a novel algorithm to achieve optimal EMS for splitting power of sources.
- Combine ECMS to refine algorithm to optimize fuel consumption and fuel cell lifespan.
- Utilize experimental data to confirm effectiveness of the proposed algorithm.

ARTICLE INFO

Keywords:
Fuel cell hybrid electric vehicle
Energy management strategy
Reinforcement learning
Data driven
Hierarchical power splitting

ABSTRACT

A reinforcement-learning-based energy management strategy is proposed in this paper for managing energy system of Fuel Cell Hybrid Electric Vehicles (FCHEV) equipped with three power sources. A hierarchical power splitting structure is employed to shrink large state-action space based on an adaptive fuzzy filter. Then, the reinforcement-learning-based algorithm using Equivalent Consumption Minimization Strategy (ECMS) is proposed for tackling high-dimensional state-action space, and finding a trade-off between global learning and real-time implementation. The power splitting policy based on experimental data is obtained by using reinforcement learning algorithm, which allows for many different driving cycles and traffic conditions. The proposed energy management strategy can achieve low computation cost, optimal fuel cell efficiency and energy consumption economy. Simulation results confirm that, compared with existing learning algorithms and optimization methods, the proposed reinforcement-learning-based energy management strategy using ECMS can achieve high computation efficiency, lower power fluctuation of fuel cell and optimal fuel economy of FCHEV.

1. Introduction

Nowadays, environmental pollution, global warming and energy concerns urge a replacement for Internal Combustion Engine (ICE) based vehicles. As a result, many kinds of new-generation environmentally friendly vehicles have been manufactured, among which the Hybrid Electric Vehicles (HEV), Fuel Cell Electric Vehicles (FCEV), and Battery Electric Vehicles (BEV) are standing in the most attractive research area. Apparently, BEV is the most promising substitute for ICE-based vehicles, while it is still platonic, on account of the immature traction battery technology and insufficient charging infrastructure [1]. To address this problem, HEV that uses both engine and motor as the hybrid power suppliers has been proposed. However, HEV still relies on fossil fuel, which means the Green House Gases (GHG) and environmental pollutants will be discharged inherently. Given this, FCEV or FCHEV without engine is proposed. Although considering many challenges are still in the air, many major vehicle manufacturers are interested in research and development of FCHEV, especially the energy

☆ This paper was supported by National Natural Science Foundation of China (Grant Nos. 61473115, U1704157), the Scientific and Technological Innovation Leaders in Central Plains (Grant No. 194200510012) and Science, Technology Innovative Teams in University of Henan Province (Grant No. 18IRTSTHNO11) and Key Scientific Research Projects of Colleges and Universities in Henan Province (Grant Nos. 19A413007, 20A120008) and National Thirteen-Five Equipment Pre-Research Foundation of China (Grant Nos. 61403120207, 61402100203).

* Corresponding author. School of Information Engineering, Henan University of Science and Technology, Luoyang, China.
** Corresponding author. School of Information Engineering, Henan University of Science and Technology, Luoyang, China.
E-mail addresses: fuzhumu@haust.edu.cn (Z. Fu), taofazhan@haust.edu.cn (F. Tao).

https://doi.org/10.1016/j.jpowsour.2020.227964
Received 28 November 2019; Received in revised form 10 February 2020; Accepted 25 February 2020
Available online 4 March 2020

management problem, which is one of the key considerations among all technologies involved in FCHEV, aiming to mitigate the environmental degradation, and improve the fuel economy and performance of power sources in real-time [2,3].

The concept of FCHEV was suggested by McElroy in 1983 [4], while the management problem of multiple power sources for FCHEV was just researched in the late 1990s, where Ultra-Capacitor (UC) bank as a peak power unit provides benefits in the light of vehicle performance, lifespan of onboard battery and energy economy [5]. During this period, almost all management strategies are rule-based. In recent decade, the development of researches on energy management strategy draws a relatively clear path: from rule-based strategies to optimization-based strategies [6]. Rule-based strategies have a relatively small computation load, but rely deeply on the experiences from experts and it is difficult to guarantee the obtainment of global optimal strategies by following the predefined rules, which lack of adaptiveness to deal with time-varying scenarios [7–9]. To deal with this problem, optimization-based approaches are emerged [10–12], which can be categorized into two groups: global optimization methods and local (real-time) optimization methods. The former, like Dynamic Programming (DP) [13], and Heuristic Dynamic Programming (HDP) [14], can explore the global optimal strategy, but require a heavy computation load, while the latter methods, like ECMS and Model Predictive Control (MPC) [15,16], need less computation load, but just having the ability to find the local optimal results.

In order to find a better trade-off technique to compromise precision and computation load, many new kinds of methodologies are presented, such as game theory [17], and Artificial Intelligence (AI) algorithms [18,19]. Inspired by improvement of Machine Learning (ML), Reinforcement Learning (RL) has been attracting attention of researchers and engineers, which has both advantages of global optimization and local optimization by learning globally and applying locally. Many results on energy management for vehicles equipped with multiple power sources [20–28] are obtained using RL algorithms. Xiong et al. proposed a Kullback-Leibler (KL) divergence based RL algorithm for plug-in HEV to renew the Transition Probability Matrix (TPM) and optimal control strategy in real time, and the simulation results indicated the proposed RL-based energy management strategy can significantly reduce the fuel consumption and can be applied in real time [26]. To upgrade the obtained optimal policy, Yuan et al. introduced a hierarchical energy management strategy for FCHEV to realize real-time application and global optimization, and proposed a new prediction model using K-Nearest Neighbor (KNN) technique to forecast the short- and long-term velocities [27]. In Ref. [28], a deep deterministic policy gradients based RL approach was applied to solve the management problem of series-parallel plug-in hybrid electric bus under a fixed driving condition, the simulation results shown that the proposed method has a great performance close to DP. Considering high-dimensional state--action spaces of Energy Management Systems (EMS), some new methods combined with RL, such as deep convolutional neural network, have been introduced [29–31]. From the aforementioned results [26–28], considering the large state-action space of EMS, most of them focus on hybrid vehicles equipped with two power sources, which is proved to be one effective method for dealing with relative simple configuration of EMS of vehicles. To our best knowledge, a few literature applies the RL technique to address the energy management problem for FCHEV equipped with Fuel Cell (FC), Battery (BAT) and UC, on account of the complex and flexible structure of EMS. Meantime, the proposed algorithms in Ref. [29,30] highly depend on computing capacity to tackle large-scale data ascribed to complexity of EMS, which may result into a huge computation load in real application.

Therefore, in this paper, an RL-based energy management strategy for FCHEV is proposed combining ECMS technique to achieve low computation cost, optimal FC efficiency and energy consumption economy. Considering the large state-action space of EMS, hierarchical structure is employed to cut the space to reduce computation load by using an adaptive fuzzy filter to separate power demand to two parts for UC and BAT/FC, respectively [38]. To improve lifespan and fuel efficiency of FC, multi-objective optimization based ECMS is suggested. Aiming to realize fast training for large-scale data and obtaining the optimal splitting strategy, a speedy learning based RL algorithm is proposed, through which the optimal splitting strategy for BAT and FC can be obtained via experimental data applying data driven technique in a relatively short period of time. To verify the validity and superiority of the proposed RL-based optimal policy, DP and an existing RL-based methods are involved in terms of computation efficiency, and finally an ECMS-based strategy is selected as a competitor considering the optimization capability with many different typical driving cycles.

The remaining part of this paper is organized as follows. The power train of FCHEV and three power sources are modeled and built in detail in Section 2. In Section 3, the hierarchical structure of RL-based energy management system with ECMS is established. In Section 4, an ECMS based Q-learning algorithm are studied and proposed, then according to historical data, the optimal policy is generated. Comparative study and analysis are carried on and discussed meticulously in Section 5. And Section 6 presents the conclusions.

2. Background and problem statement

In this section, the considered vehicle model is constructed firstly based on the platform as shown in Fig. 1. Then, the detailed mathematical models of three power sources are introduced.

2.1. System configuration and structure of FCHEV

With the rapid development of the technology on vehicular power system, many kinds of propulsion systems with different topological structures have been designed aiming to certain purposes [32]. Here, according to the test bed, detailed configurations of power train and EMS of the investigated FCHEV are clearly shown in Fig. 2.

In this configuration, the directly controlled objects comprise DC/DC converters, and DC/AC inverters. By manipulating these objects, the output power of hydrogen FC, li-ion BAT, and UC can be regulated to match up the power demand. Furthermore, the vehicle has a three-phase traction motor, and a DC bus for power sharing. The specific parameters of involved FCHEV are listed in Table 1.

The primary power source for FCHEV is FC. Unidirectional DC/DC converter serves as an intermediate layer for linking FC to DC bus, and as a regulator for maintaining the State of Charge (SoC) of BAT pack at a proper level, on the premise that FC is working in high-efficiency field. By applying the bi-directional DC/DC converter, UC tunes the DC bus power, and can produce or retrieve peak power, as the vehicle's instantaneously strong acceleration/deceleration, and BAT pack has ability to provide or absorb the rest power through DC bus. Additionally, DC/AC inverter generates any desired power for traction motor to drive the vehicle for meeting the driver's demand.

2.2. Modeling of power sources

The major aim of this research is to find a new approach to reduce the computation load to realize global optimization and real-time learning for multi-source FCHEV with complex structure of energy system and large-scale experimental data. Due to the energy management of FCHEV is a typical multi-objective optimization problem, ECMS is considered to realize real-time learning process, and achieve the highest FC operational efficiency as well as the longest lifespan of UC and BAT. In ECMS, optimization objectives are fuel consumption and lifespan of power sources, involving FC, BAT and UC, which can be modeled as follows:

(1) Fuel cell model

The Proton-Exchange-Membrane Fuel-Cell (PEMFC) as the major

Fig. 1. Structural configuration model of vehicle test bench.

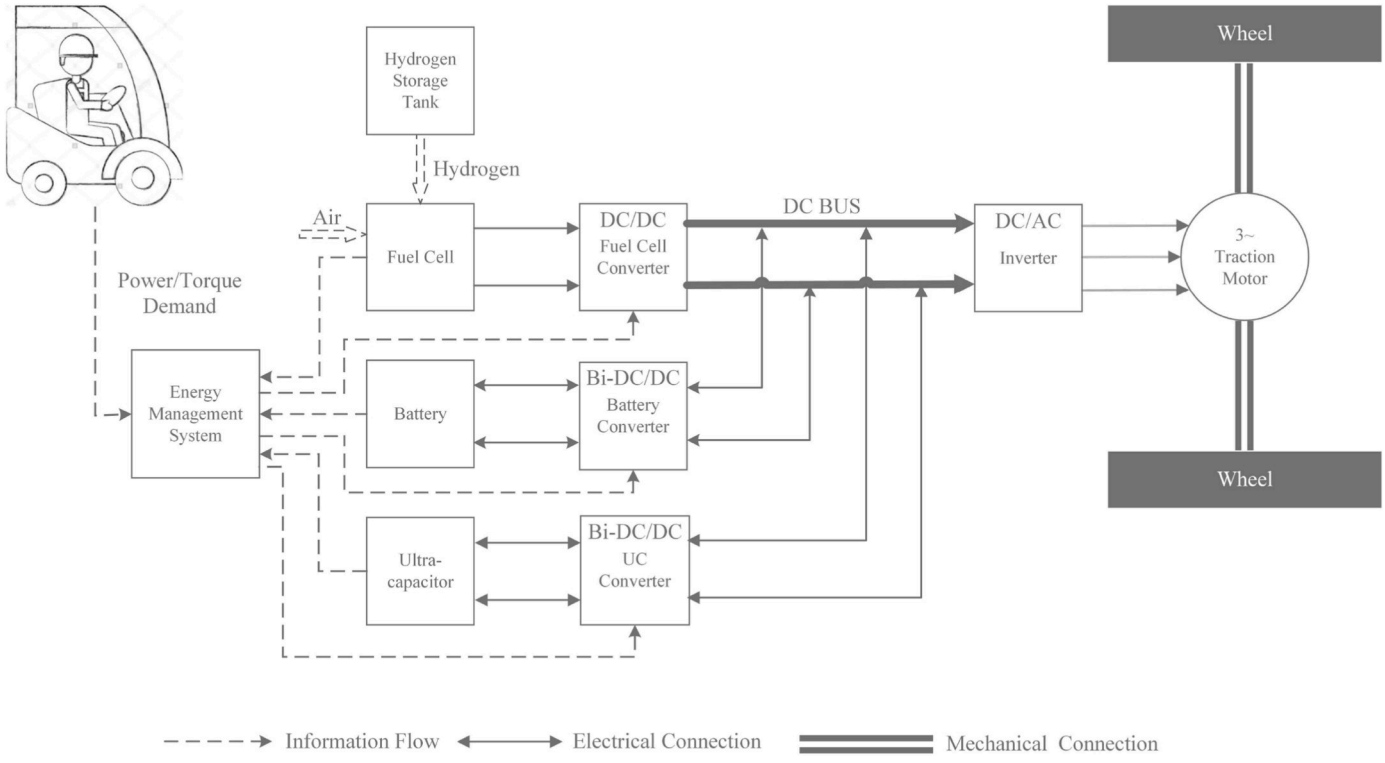


Fig. 2. Block diagram of energy management system in multi-source FCHEV.

Table 1
Main parameters of the considered hydrogen FCHEV.

Device	Item	Parameter
Electric motor	Power rating	45 kW
	Rated speed	1500 rpm
Fuel cell	Power rating	10 kW
	Output voltage	40–100V
Ultracapacitor	Rated voltage	288V
	Capacitance	27.5F
	Energy	320 Wh
Battery	Type	Li-ion
	Energy	25.6 kWh
	Rated voltage	320V
	Maximum charge rate	2C
	Maximum discharge rate	4C
	Nominal charge/discharge rate	0.5C
DC bus	Rated voltage	AC380V

power source for FCHEV converts the chemical energy into electric energy through the reaction between hydrogen and oxygen. The output voltage of FC is expressed as [33]:

$$V_{FC} = n_{cell} \times (E_{cell} - V_{act.loss} - V_{ohm.loss}) \tag{1}$$

where n_{cell} means the number of cells on the stack, E_{cell} is the electromotive potential in volts, $V_{act.loss}$ is the losses due to the cell activation in volts, and $V_{ohm.loss}$ is the losses due to the cell internal resistance in volts.

The electromotive potential E_{cell} can be calculated by

$$E_{cell} = 1.229 - 0.85e^{-3}(T - T_c) + \frac{RT}{2F}\ln\left(\sqrt{P_{O_2}}P_{H_2}\right) \tag{2}$$

where T is the temperature of the catalyst layer in degree kelvin, $T_c = 298.15K$ is the temperature offset in degree of kelvin, $R = 8.314J\cdot(mol\cdot K)^{-1}$ is the gas constant in joules per mole degree kelvin, $F = 96485C\cdot mol^{-1}$ is the Faraday constant in coulombs per mole, P_{O_2} and P_{H_2} are the pressures at the interface of cathode and anode catalyst layer in pascals, respectively.

The activation losses $V_{act.loss}$ can be expressed as

$$V_{act.loss} = \frac{RT}{2\alpha F}\ln\frac{I_{FC}}{I_0 S_{cata}} \tag{3}$$

where α is charge transfer coefficient, I_{FC} is FC current in amperes, S_{cata} is the catalyst layer section area in square centimeters, and I_0 is the exchange current density in amperes per square centimeter.

The ohmic losses $V_{ohm.loss}$ can be described as

$$V_{ohm.loss} = \frac{I_{FC}\int_0^l \Gamma(T_{mem}, \lambda(z))dz}{S_{mem}} \tag{4}$$

where l is thickness of the membrane in centimeters, S_{mem} is the membrane surface area in square centimeters, and $\Gamma(T_{mem}, \lambda(z))$ is the local resistivity of the membrane in ohms centimeter, which can be obtained by

$$\Gamma(T_{mem}, \lambda(z)) = \begin{cases} \dfrac{10^3}{1.933}e^{\left[1268\cdot\left(\frac{1}{T_{mem}}-\frac{1}{303}\right)\right]} & ,0 < \lambda \le 1 \\ \dfrac{10^3}{5.193\lambda - 3.26}e^{\left[1268\cdot\left(\frac{1}{T_{mem}}-\frac{1}{303}\right)\right]} & ,\lambda > 1 \end{cases} \tag{5}$$

where T_{mem} is the temperature of the membrane in degree kelvin, and $\lambda(z)$, $z \in [0, l]$, is the water content of the membrane.

(2) Battery model

The most classical method to estimate SoC of BAT is current integration [34], which can be expressed as

$$SoC_{BAT} = SoC_{BAT.ini} + \frac{\beta\int i_{BAT}dt}{C_{nom}} \tag{6}$$

where SoC_{ini} is the initial BAT SoC, i_{BAT} is the BAT current in amperes, $\beta = \pm 1$ is a charge-discharge switch (positive to the charge and negative during discharge) and C_{nom} is the nominal BAT capacity in ampere hour.

The output voltage V_{BAT} is attained by

$$\begin{cases} V_{BAT} = V(SoC_{BAT})_{BAT.oc} + \beta i_{BAT} r(SoC_{BAT}) \\ i_{BAT} = \dfrac{V(SoC_{BAT})_{BAT.oc} - \sqrt{V^2(SoC_{BAT})_{BAT.oc} - 4r(SoC_{BAT})P_{BAT}}}{2r(SoC_{BAT})} \end{cases} \tag{7}$$

where $V(SoC_{BAT})_{BAT.oc}$ means the open circuit voltage at the state of charge SoC_{BAT} of BAT in volts, $r(SoC_{BAT})$ is the internal resistance at the state of charge SoC_{BAT} in ohms, and P_{BAT} means the electric power of BAT in watts.

(3) Ultracapacitor model

The open circuit voltage of UC $V_{UC.oc}$, the SoC of UC SoC_{UC} and the UC current I_{UC} can be formulated as [35].

$$V_{UC.oc} = SoC_{UC}\cdot(V_{UC.\max} - V_{UC.\min}) + V_{UC.\min} \tag{8}$$

$$I_{UC} = \frac{V_{UC.oc} - \sqrt{V_{UC.oc}^2 - 4R_{UC}P_{UC}}}{2R_{UC}} \tag{9}$$

where $V_{UC.max}$ and $V_{UC.min}$ are the maximum and minimum voltage of UC in volts, respectively, R_{UC} is equivalent internal resistance in ohms, and P_{UC} is electric power of UC in watts.

3. Hierarchical structure of RL-based EMS

In this section, in view of the large learning space, a hierarchical structure for power split in EMS is presented to shrink the state-action space to short computation time of learning historical data. First, detailed configuration of the hierarchical EMS is pictured in Fig. 3. Then, additionally, the adaptive fuzzy filter and ECMS embedded in EMS are discussed and modeled in detail.

3.1. Fuzzy based adaptive low-pass filter

The purpose of applying low-pass filter is to protect FC and BAT from power fluctuations that UC provides/absorbs the peak power to upgrade the power performance of vehicles. The low-pass filter can be formulated as

$$G(s) = \frac{1}{\frac{1}{f_s}s + 1} \tag{10}$$

where f_s is the regulating frequency, tuned by a Fuzzy Inference System (FIS) according to P_{demand}, SoC_{UC}, and SoC_{BAT}.

Here a compound parameter $SoC_{ESS} = \alpha_{UC}SoC_{UC} + \alpha_{BAT}SoC_{BAT}$, where α_{UC} and α_{BAT} are adjustment coefficients, is presented to describe a comprehensive SoC of ESS to guarantee SoC_{UC} and SoC_{BAT} vary in a given range, and to alleviate FC's working load. SoC_{ESS} and P_{demand} are selected as the input variables of FIS, and output is served by f_s. Through trial and error, a proper fuzzy rule base is built on experiences shown in Table 2.

3.2. ECMS based on multi-objective optimization theory

In FCHEVs, ECMS can be applied to manage power flow and upgrade

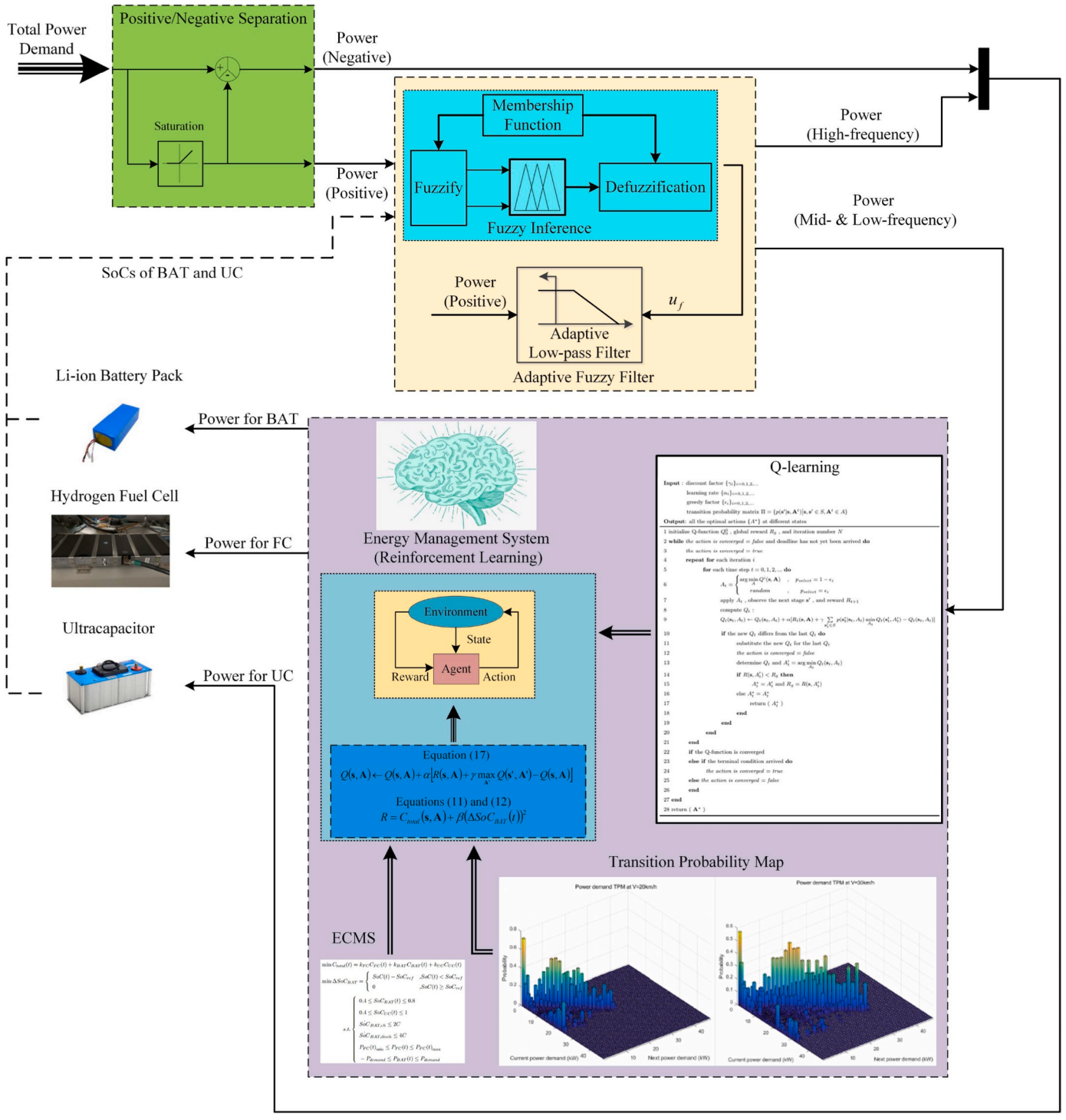


Fig. 3. Optimized power splitting structure based on reinforcement learning.

fuel economy locally, when extra energy storage sources like BAT and UC are introduced to offer some supportive helps to FC for meeting the driver's power demand [3]. According to the detailed expressions of each power source, in terms of the total hydrogen fuel consumption and the SoC deviation of BAT, the ECMS can be expressed as

$$\begin{aligned} &\min C_{total}(t) = k_{FC}C_{FC}(t) + k_{BAT}C_{BAT}(t) + k_{UC}C_{UC}(t) \\ &\min \Delta SoC_{BAT} = \begin{cases} SoC(t) - SoC_{ref} & , \quad SoC(t) < SoC_{ref} \\ 0 & , \quad SoC(t) \geq SoC_{ref} \end{cases} \end{aligned} \tag{11}$$

$$\begin{cases} 0.4 \leq SoC_{BAT}(t) \leq 0.8 \\ 0.4 \leq SoC_{UC}(t) \leq 1 \\ s.t. \; \dot{SoC}_{BAT,ch} \leq 2C \\ \dot{SoC}_{BAT,disch} \leq 4C \\ P_{FC}(t)_{\min} \leq P_{FC}(t) \leq P_{FC}(t)_{\max} \\ -P_{demand} \leq P_{BAT}(t) \leq P_{demand} \end{cases} \tag{12}$$

where $C_{total}(t)$ is total volume of hydrogen consumption in grams at time t, $C_{FC}(t)$ is FC hydrogen consumption in grams at time t, $C_{BAT}(t)$ and $C_{UC}(t)$ are BAT and UC equivalent hydrogen consumption in grams at time t, respectively. k_{FC} means the FC efficiency penalty coefficient that

Table 2
Fuzzy rule base for FIS.

f_s	P_{demand}							
	NB	NM	NS	ZE	PS	PM	PB	NB
SoC_{ESS}	S	S	S	RS	B	RB	M	RS
	RS	S	RS	MS	B	RB	M	RS
	M	RS	RS	M	B	M	M	S
	RB	M	M	RB	B	RS	RS	S
	B	RB	RB	B	B	RS	S	S

N=Negative, P=Positive, S=Small, M = Medium, B=Big, R = Relatively, ZE = Zero.

allows FC operating at high efficiency level, k_{BAT} and k_{UC} are penalty coefficient in terms of the SoC value of BAT and UC, respectively. $\Delta SoC_{BAT}(t)$ is deviation of the current SoC of BAT from the reference value at t. According to the physical properties of each power source on the platform, some boundaries of parameters are given in equation (12), where $P_{FC}(t)_{min}$ and $P_{FC}(t)_{max}$ are the minimum and maximum output power, respectively, as the FC works in the high-efficiency field. All the boundary values of constraints are determined by the real vehicle test bench.

The FC hydrogen consumption can be obtained by

$$C_{FC}(t) = \int_0^t \left(\frac{1.2 M_{H_2} N_{cell}}{2F} I_{FC}(t) \right) dt \tag{13}$$

where $M_{H_2} = 2g{\cdot}mol^{-1}$ represents the hydrogen molar mass, N_{cell} means the number of cells, $F = 96487C{\cdot}mol^{-1}$ is the Faraday constant, and $I_{FC}(t)$ is the fuel cell current in amperes.

The BAT equivalent hydrogen consumption can be expressed as

$$C_{BAT}(t) = \begin{cases} \dfrac{P_{BAT}{\cdot}C_{FC,ave}}{\eta_{disch}{\cdot}\eta_{ch,ave}{\cdot}P_{FC,ave}} & , P_{BAT} \geq 0 \\ \dfrac{P_{BAT}{\cdot}\eta_{disch,ave}{\cdot}\eta_{ch}{\cdot}C_{FC,ave}}{P_{FC,ave}} & , P_{BAT} < 0 \end{cases} \tag{14}$$

where $P_{BAT}(t)$ is the BAT power at time t in watts, $C_{FC,ave}$ is the average hydrogen consumption of FC, $P_{FC,ave}$ represents the mean power of FC, η_{disch} and η_{ch} means the discharging and charging efficiency, $\eta_{disch,ave}$ and $\eta_{ch,ave}$ means the average discharging and charging efficiency. $P_{BAT}(t) \geq 0$ means the BAT is discharging, and $P_{BAT}(t) < 0$ means charging.

The equivalent hydrogen consumption of UC can be calculated same to BAT, where k_{FC} , k_{BAT} and k_{UC} are chosen followed by Ref. [3].

4. Learning for Markov Chain models and Q-learning algorithm

In this section, the power demand varying with vehicle velocity is regarded as a finite-state Markov Chain (MC) problem, and Q-learning is used to solve the Markov Decision Process (MDP).

4.1. Q-learning for MDP

MDP framework is considered with using Q-learning technique, whose aim is to gain an optimal policy π^* that maximizes the expected discounted long-term reward

$$V^*(\mathbf{s}) = \max_{\pi} E\left[\sum_{t=0}^{\infty} \gamma^t R(\mathbf{s}^t, \pi(\mathbf{s}^t)) \middle| \pi, \mathbf{s}^0 = \mathbf{s} \right]$$

for each state $\mathbf{s}$, where $\gamma \in [0,1)$ is the discount factor. Q-functions represent the expected future discounted reward for a state $\mathbf{s}$ when an action $\mathbf{A}$ is performed. The optimal Q-function Q^* satisfies the Bellman equation:

$$Q^*(\mathbf{s},\mathbf{A}) = R(\mathbf{s},\mathbf{A}) + \gamma \sum_{\mathbf{s'}} [p(\mathbf{s'}|\mathbf{s},\mathbf{A}) \max_{\mathbf{A'}} Q^*(\mathbf{s'},\mathbf{A'})] \tag{15}$$

Here, it assumes the environment is stationary. Hence, the superscript t is ignored, and consequently, $\mathbf{s}$ is used to represent the current state $\mathbf{s}^t$, and $\mathbf{s}$' is used to substitute the next state $\mathbf{s}^{t+1}$.

Considering the computation load, to determine a proper state-action space, UC is only involved to provide or absorb sudden power demand and negative power, and then FC and BAT work together for the rest. Therefore, in this paper, MDP consists of a set of actions $A = \{P_{FC}, P_{BAT}\}$, a set of state variables $S = \{SoC_{BAT}(t), SoC_{UC}(t), P_{demand}(t), V_{current}(t)\}$, and a reward function $R = C_{total}(\mathbf{s},\mathbf{A}) + \beta(\Delta SoC_{BAT}(t))^2$, when β is a positive penalty coefficient, and $\Delta SoC_{BAT}(t)$ is the same term as presented in equation (11).

To calculate equation (15) and get the optimal policy, it is critical to properly estimate $Q^*(\mathbf{s},\mathbf{A})$. Q-learning is a widely used model-free off-policy learning approach and starts with an initial $Q(\mathbf{s},\mathbf{A})$ for each state-action pair [36]. At each time step, all agents perform a joint action based on a commonly used exploration method ε-greedy strategy that selects the greedy action $\arg\max_{\mathbf{A}} Q(\mathbf{s},\mathbf{A})$ with high probability, and betweenwhiles, selects an action uniformly at random with a small probability ε. Each time a joint action $\mathbf{A}$ is taken in state $\mathbf{s}$, then the reward $R(\mathbf{s},\mathbf{A})$ is fed back from environment and the next state $\mathbf{s}$' is observed, thus the Q-value is updated with a combination of its current value and the Temporal-Difference Error (TDE), expressed as

$$Q(\mathbf{s},\mathbf{A}) \leftarrow (1-\alpha) Q(\mathbf{s},\mathbf{A}) + \alpha[R(\mathbf{s},\mathbf{A}) + \gamma \max_{\mathbf{A'}} Q(\mathbf{s'},\mathbf{A'})] \tag{16}$$

which can be rewritten as

$$Q(\mathbf{s},\mathbf{A}) \leftarrow Q(\mathbf{s},\mathbf{A}) + \alpha[R(\mathbf{s},\mathbf{A}) + \gamma \max_{\mathbf{A'}} Q(\mathbf{s'},\mathbf{A'}) - Q(\mathbf{s},\mathbf{A})] \tag{17}$$

where $\alpha \in (0,1)$ means the learning rate which reflects the influence of the new experience on current estimation $Q(\mathbf{s},\mathbf{A})$.

4.2. Improved learning algorithm based on Q-learning

The configuration of FCHEV equipped with UC, BAT, and FC is complex due to the coupled output power and flexible topology of power system, which is different from the other types of vehicles, like the ICE vehicles. To achieve a lower hydrogen consumption and a better power performance, it is necessary to describe vehicles more comprehensively, resulting in a larger state space, but a higher computation load. Conventional Q-learning based technique can successfully manage the power split under the less-state less-action conditions. Nevertheless, confronted with high-dimensional state-action space, or even continuous state-action variables, it is difficult to be solved efficiently by existing RL algorithms, as a result of the "Curse of Dimensionality", although some more powerful algorithms are presented, like deep learning [37], which inevitably requires a powerful computing performance. Hence a fast learning algorithm based on Q-learning and ECMS is designed and developed to tackle this irksome problem and increase the convergence speed without degrading the optimality of results. Pseudo code is shown in Table 3.

4.3. Data-driven optimal EMS based on improved Q-learning

By applying the proposed algorithm, the experimental data can be learned to get TPM of power demand for optimizing power split of EMS. In this research, to learn a better management strategy, about 150 thousand sets of experimental driving data (the years from 2017 to 2019) are collected [39–41]. After removing the invalid data, more than 40 thousand data points are selected.

To get the TPM of power demand, maximum likelihood estimation and nearest neighbor method are considered, expressed as

$$p_{k,ij} = \frac{N_{k,ij}}{N_{k,i}}$$

Table 3
Pseudo code of the proposed QL algorithm.

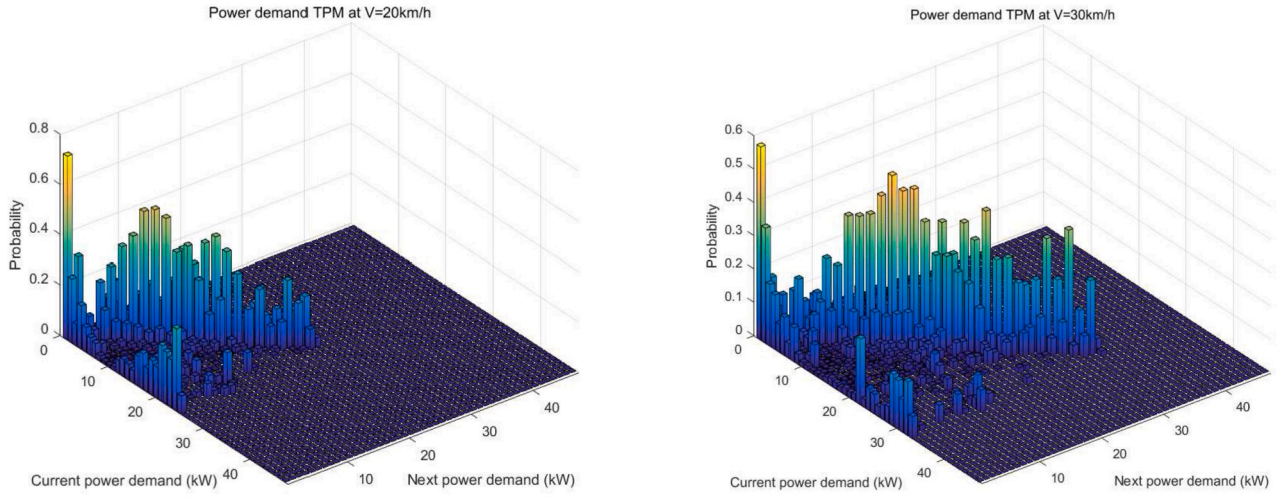
Pseudo code of the speedy-based Q-learning algorithm for FCHEV EMS

Input : discount factor $\{\gamma_i\}_{i=0,1,2,...}$
learning rate $\{\alpha_i\}_{i=0,1,2,...}$
greedy factor $\{\epsilon_i\}_{i=0,1,2,...}$
transition probability matrix $\Pi = \{p(\mathbf{s}'|\mathbf{s}, \mathbf{A}^t)|\mathbf{s}, \mathbf{s}' \in S, \mathbf{A}^t \in A\}$
Output: all the optimal actions $\{A^*\}$ at different states

initialize Q-function Q_0^0 , global reward R_g , and iteration number N
while *the action is converged* = *false* and deadline has not yet been arrived **do**
  the action is converged = *true*
  repeat for each iteration i
    for each time step $t = 0, 1, 2, ...$ **do**
      $A_t = \begin{cases} \arg\min_A Q^i(\mathbf{s}, \mathbf{A}) & , \quad p_{select} = 1 - \epsilon_i \\ random & , \quad p_{select} = \epsilon_i \end{cases}$
      apply A_t , observe the next stage $\mathbf{s}'$, and reward R_{t+1}
      compute Q_t :
      $Q_t(\mathbf{s}_t, A_t) \leftarrow Q_t(\mathbf{s}_t, A_t) + \alpha[R_t(\mathbf{s}, \mathbf{A}) + \gamma \sum_{\mathbf{s}'_t \in S} p(\mathbf{s}'_t|\mathbf{s}_t, A_t) \min_{A_t} Q_t(\mathbf{s}'_t, A'_t) - Q_t(\mathbf{s}_t, A_t)]$
      if the new Q_t differs from the last Q_t **do**
        substitute the new Q_t for the last Q_t
        the action is converged = *false*
        determine Q_t and $A'_t = \arg\min_{A_t} Q_t(\mathbf{s}_t, A_t)$
        if $R(\mathbf{s}, A'_t) < R_g$ **then**
          $A^*_t = A'_t$ and $R_g = R(\mathbf{s}, A'_t)$
        else $A^*_t = A^*_t$
          return (A^*_t)
        end
      end
    end
  end
  if the Q-function is converged
  else if the terminal condition arrived **do**
    the action is converged = *true*
  else *the action is converged* = *false*
  end
end
return ($\mathbf{A}^*$)

where $N_{k,ij}$ is the frequency of occurrence that the power demand P_{demand} transits from P^i_{demand} to P^j_{demand} at a certain vehicle velocity of v_k, and $N_{k,i}$ means the total counts of the frequency of occurrence that the power demand P_{demand} transits from P^i_{demand} to all possible power demands at a certain vehicle velocity of v_k.

For the sake of simplification of data processing, the selected data are classified into 36 groups, then it is easy to get the TPM at different velocities varying from 0 km/h to 50 km/h in 10 km/h increments. The average TPM at each level of velocity can be calculated. Fig. 4 shows the



(a) Power demand transition probability at the velocity of 20 km/h **(b)** Power demand transition probability at the velocity of 30 km/h

Fig. 4. Transition probability maps of power demand at different velocities.

average transition probability maps at velocities of 20 km/h and 30 km/h.

According to the proposed algorithm, by using the obtained TPM, it is easy to get the optimal management strategy, shown in Fig. 5 (taking the scenario $V = 20$ km/h and the SoCs of BAT and UC are all equal to 0.7 as an example). The computation time of different optimization methods is listed and compared in Table 4 that shows the learning efficiency of the proposed method.

Table 4
Computation time comparison for training by three optimization methods.

Optimization method	Computation time[a](hours)	Improvement(%)
DP	164	–
RL	97	40.9
Proposed RL-based method	41	57.7

[a] A 2.5 GHz microprocessor with 4 GB RAM was used.

5. Simulation and analysis

In this section, the optimal energy management strategy learned by RL-based method is applied to several commonly used driving cycles to testify the availability and validity of the proposed algorithm and rationality of using RL technique to solve the energy management problem for three-source FCHEV.

To verify the universality of proposed optimal strategy, an experimental driving cycle is discussed first, besides which six typical commonly used driving cycles are considered. They are HWFET, UDDS, WVUCITY, WVUINTER, WVUSUB, and a compound driving cycle imitating a vehicle runs from a city to another city through two arterial roads and an expressway. The simulation results are shown in Figs. 6 and 7, as well as Supplementary Figs. 1–5 (Supplementary Figs. 1–5 are shown in Supplementary document for details).

According to the simulation results shown in Fig. 6, it can be known that the peak power is provided/absorbed by UC, and during the driving cycle, the output power of FC is high enough and relatively stable resulting in the operational efficiency of FC is located in high-efficiency field, and the SoC value of BAT decreases mildly, except that as the vehicle runs at about 1300s, where exists a quick and strong acceleration needing much more sudden power than before, so all power sources have to provide more power to support the rapid demand, hence, at the same time, the SoCs of UC and BAT are all reduced, and the efficiency of FC is fluctuated.

The driving cycles presented in Supplementary Figs. 1 and 2, can be categorized into a group, which is used for emission certification and fuel economy testing of light-duty vehicles. Analyzing the results of these two normal driving cycles, some common characteristics can be found: as the vehicle speeds up, the output power and operational efficiency of FC grow up as well, namely, the variation tendency of output power from FC approximates that of vehicle velocity, and thereupon the variation affects the efficiency of FC; and if velocity is relatively stable, UC and BAT hardly do any efforts resulting in a stable SoC value.

By studying the three figures above, it is clearly recognized for light-duty vehicles, under the three testing driving cycles, the obtained optimal policy presents an ideal performance. After using the proposed energy management strategy, the power provided from FC is relatively high and stable guaranteeing the high efficiency (greater than 0.5 almost at any time) and long lifespan; the power from BAT is fluctuated in an acceptable variation range ensuring the lifespan of it as well (the energy consumption of BAT is about 5% per 800 s no matter what kind of the driving cycle is seem to be). And the SoCs of BAT and UC are reduced to some extents, which lead to the lower fuel consumption, higher fuel economy and longer driving range. Two exceptional things should be notated are, in Supplementary Fig. 1, firstly, at the end of the driving cycle, there exists a strong and rapid deceleration inducing the UC absorbs a mass of braking energy resulting in the abruptly increasing SoC value in a short time, and, secondly, in Supplementary Fig. 2, due to a great deal of the times of rapid deceleration, the tendency of SoC of UC is generally increasing leading to a relatively low efficiency of FC, although the efficiency is still located in high-efficiency field (greater than 0.5).

To verify the universality of proposed strategy, besides the driving cycles studied above, some more severe driving conditions are

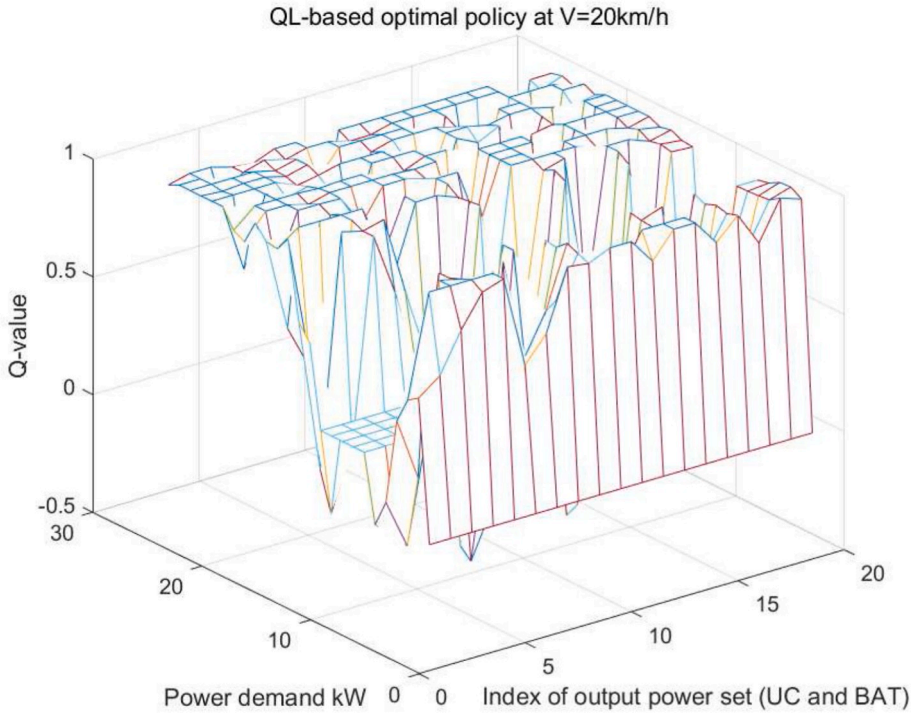


Fig. 5. QL-based optimal policy under the condition of $V = 20$ km/h, and $SoC_{BAT} = SoC_{UC} = 0.7$

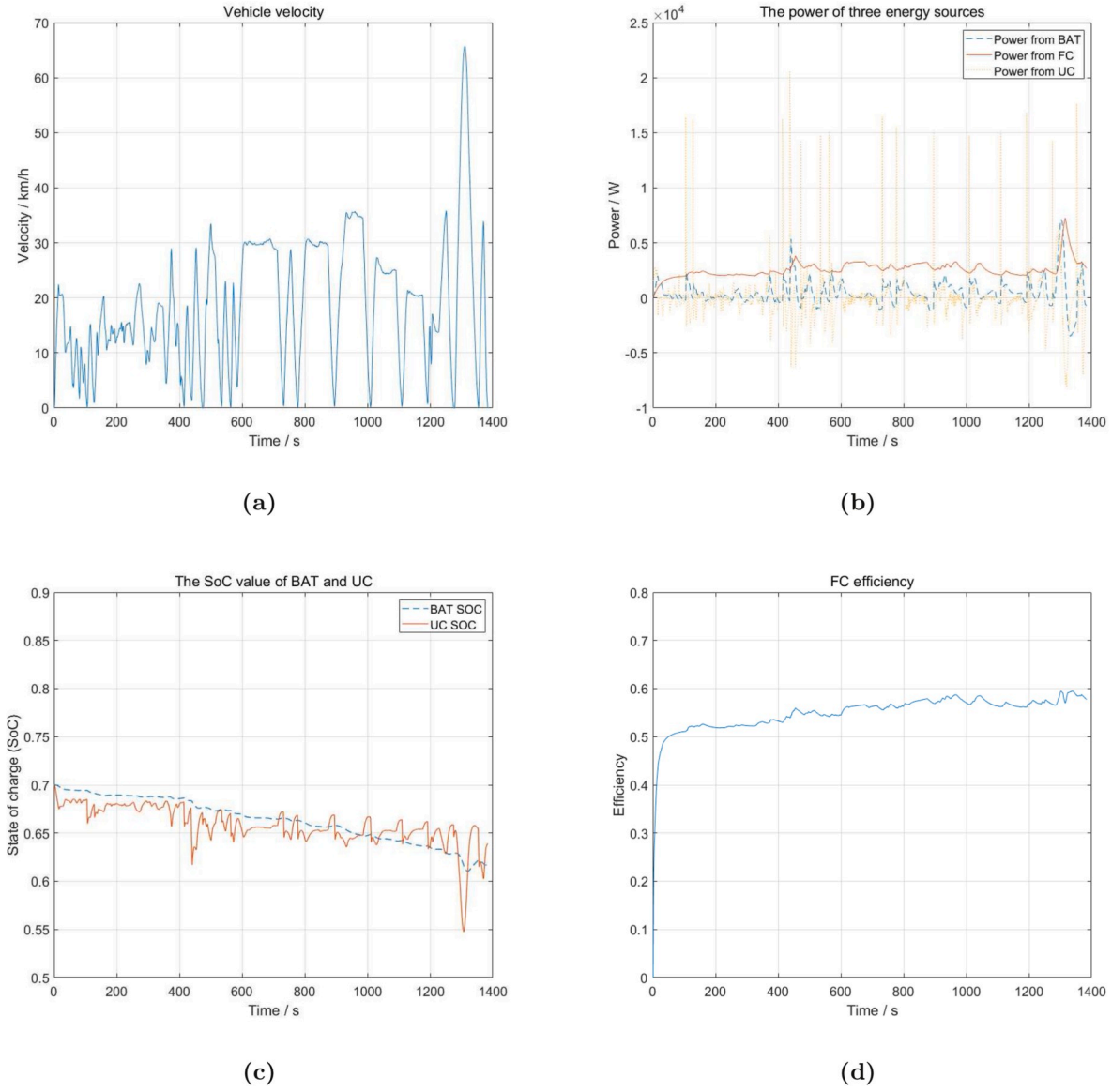


Fig. 6. Simulation results of an experimental driving cycle.

taken into consideration.

From Supplementary Figs. 3–5, it is reasonably seen that there are some apparent differences between Supplementary Figs. 1–2 and Supplementary Figs. 3–5. In the latter group, it is obvious the driving cycles are more uncertain, nevertheless, the output power and SoCs of three power sources perform well same as the former group. Something more important is that the efficiency of FC is degrading, especially in driving cycle WVUCITY, owing to the extreme driving conditions, the use ratio of UC is very high as well as the maximal output power of it, therefore, together with the low average velocity, the efficiency of FC is only varying in the range from 0.4 to 0.5. With regard to the driving cycle WVUINTER, thanks to high average velocity, particularly during the time interval from 200 to 1300 s, the efficiency of FC is located in high-efficiency field, and after this interval, the speed is getting lower resulting in a descending fuel efficiency. For cycle WVUSUB, from the analysis above, it is logically seen in Supplementary Fig. 5 that UC is frequently used to absorb sharply braking energy, which brings about the rising of its SoC value, and the frequent strong acceleration/deceleration behaviors make the fuel efficiency fluctuating and relatively low (because of the higher speed compared with WVUCITY, the fuel efficiency of WVUSUB is,consequently, higher than that of WVUCITY).

In each discussion above, it consists of only one single driving cycle, which is just thinking about the short-trip scenario. So here a kind of long-trip scenario simulates a travel that the driver wants to drive his/her vehicle from the downtown he/she lives in to another downtown he/she aims to arrive to. The whole driving cycle is illustrated in Fig. 7(a).

From Fig. 7(b) and (c), some common properties same as the previous discussions can be found, additionally, some new findings can be discovered. As the vehicle is driven on the arterial roads, the steady variation of velocity and the relatively high driving speed make the SoC of UC grow up, and make the efficiency of FC stands in the high-efficiency field. Furthermore, as shown in Fig. 7(d), the result, which is more important and not be mentioned before, is if the vehicle runs after the high-speed phases (like, arterial road or expressway), the fuel efficiency and the SoC of UC will be elevated to a new higher level accordingly within a certain period of time. This finding may have the ability to guide the optimal distribution of hydrogen refueling stations alongside the expressway in the future.

According to the discussions above, it is reasonably recognized that the proposed energy management strategy based on RL technique can manage these different power sources for meeting the power demand from the driver coordinatively with a harmonious relationship. And by

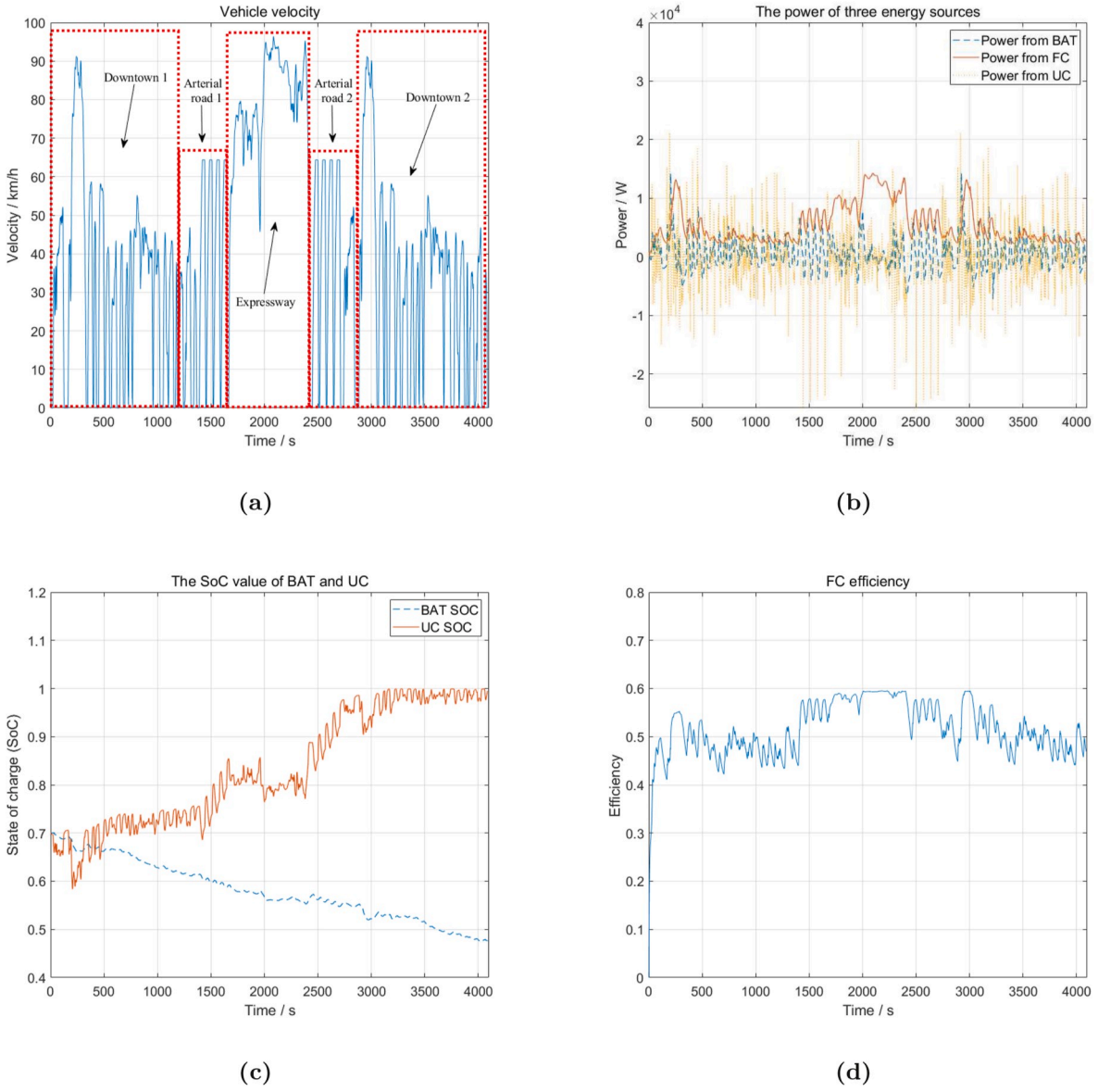


Fig. 7. Simulation results of a compound driving cycle.

using this strategy, the lifespan of BAT and FC can be extended. To testify the better performance on fuel consumption of the proposed policy, a simple comparison is made, listed in Table 5. In this comparison, the RL-based results are gained by using proposed RL algorithm considering ECMS, while the dissimilitude is, in this paper, the ECMS-based ones are calculated by conventional single objective optimization based ECMS.

As the table shows, the RL-based strategy improves the fuel economy further, compared with the traditional ECMS-based strategy. In particular, the RL-based strategy performs better if the vehicle runs in an urban area, has a long-trip, and be driven by a reckless driver with bad driving behaviors, due to the learning ability of RL technique that can deal with many unexpected conditions, which is not possessed by ECMS-based one.

In summary, according to Table 4, it is clearly recognized the proposed RL-based method has the lowest computation complexity compared with DP and conventional RL. And as shown in Table 5, the comparison table shows that the proposed RL-based strategy can lead to a lower hydrogen consumption confronted with traditional ECMS-based strategy, which has been proved to achieve near-optimal performance [42]. And as use the proposed RL-based strategy to settle many different typical driving cycles, the equivalent hydrogen consumptions of all driving cycles are lower than those obtained by traditional ECMS-based strategy, which means the proposed RL-based strategy has a better

Table 5
The comparison of RL- and ECMS-based strategies on equivalent fuel consumption.

Driving cycle	Equivalent fuel consumption (L/100 km)		Improvement
	RL-based strategy	ECMS-based strategy	
Experimental driving cycle	3.6	4.4	18.2%
HWFET	2.6	2.8	7.1%
UDDS	2.9	3.2	9.4%
WVUCITY	5.0	6.1	18.0%
WVUINTER	2.9	3.2	9.4%
WVUSUB	3.2	3.5	8.6%
Compound driving cycle	2.8	3.8	26.3%

feasibility compared with the traditional ECMS-based one.

6. Conclusions

In this paper, RL technique was used to solve the energy management problem for FCHEV. First, according to the real test bench, the structure of FCHEV has been determined, and the detailed model of each power source has been established, based on which a multi-objective optimization problem was built according to ECMS. Then, MC model and Q-learning algorithm were studied, and a more efficient Q-learning algorithm was proposed. After that, the hierarchical structure of EMS was presented to reduce the space scale by applying an adaptive fuzzy filter. Afterwards, based on the experimental data, TPM and reward matrix were figured out, and the optimal power splitting policy was found via MATLAB. To compare the computation efficiency of the proposed optimization method, DP and conventional RL were selected. Comparative results presented that the proposed RL algorithm with a hierarchical power splitting structure has an ability to reduce the computation load effectively. Finally, some commonly used driving cycles were employed to ensure the effectiveness and availability of the proposed method. Simulation results showed that the proposed RL-based optimal strategy has a more environmentally friendly and promising ability to extend the lifespan of FC and BAT, and improve the operational efficiency of FC in different driving cycles. While the use rate of UC and BAT is not high enough, on account of the using of hierarchical structure. Additionally, from this research, some energy-saving tips can be found that are driving faster without frequent strong acceleration and deceleration, and after a high-average-velocity driving period, the fuel consumption will stay in a relatively low level in a certain period of time, which will helpful, in future, for optimizing the best distribution of fuel stations.

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.

CRediT authorship contribution statement

Haochen Sun: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Zhumu Fu:** Resources, Supervision, Funding acquisition. **Fazhan Tao:** Writing - original draft, Writing - review & editing, Supervision. **Longlong Zhu:** Software, Validation, Investigation. **Pengju Si:** Writing - original draft.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpowsour.2020.227964.

References

[1] Z. Liu, H. Hao, X. Cheng, et al., Critical issues of energy efficient and new energy vehicles development in China, Energy Pol. 115 (2018) 92–97.
[2] B.G. Pollet, I. Staffell, J.L. Shang, Current status of hybrid, battery and fuel cell electric vehicles: from electrochemistry to market prospects, Electrochim. Acta 84 (2012) 235–249.
[3] Y. Hames, K. Kaya, E. Baltacioglu, et al., Analysis of the control strategies for fuel saving in the hydrogen fuel cell vehicles, Int. J. Hydrogen Energy 43 (23) (2018) 10810–10821.
[4] J.F. Mcelroy, Fuel cell power plants for automotive applications, IEEE Trans. Veh. Technol. 32 (1) (1983) 33–41.
[5] E. Faggioli, P. Rena, V. Danel, et al., Supercapacitors for the energy management of electric vehicles, J. Power Sources 84 (2) (1999) 261–269.
[6] Z. Yi, H. Liu, G. Qiang, Varying-domain optimal management strategy for parallel hybrid electric vehicles, IEEE Trans. Veh. Technol. 63 (2) (2014) 603–616.
[7] S. Njoya Motapon, L.A. Dessaint, K. Al-Haddad, A comparative study of energy management schemes for a fuel-cell hybrid emergency power system of more-electric aircraft, IEEE Trans. Ind. Electron. 61 (3) (2014) 1320–1334.
[8] M. Zandi, A. Payman, J.P. Martin, et al., Energy management of a fuel cell/supercapacitor/battery power source for electric vehicular applications, IEEE Trans. Veh. Technol. 60 (2) (2011) 433–443.
[9] Z. Song, H. Hofmann, J. Li, et al., Energy management strategies comparison for electric vehicles with hybrid energy storage system, Appl. Energy 134 (2014) 321–331.
[10] L. Xu, M. Ouyang, J. Li, et al., Application of pontryagin's minimal principle to the energy management strategy of plugin fuel cell electric vehicles, Int. J. Hydrogen Energy 38 (24) (2013) 10104–10115.
[11] L. Xu, F. Yang, J. Li, et al., Real time optimal energy management strategy targeting at minimizing daily operation cost for a plug-in fuel cell city bus, Int. J. Hydrogen Energy 37 (20) (2012) 15380–15392.
[12] Z. Yu, D. Zinger, A. Bose, An innovative optimal power allocation strategy for fuel cell, battery and supercapacitor hybrid electric vehicle, J. Power Sources 196 (4) (2011) 2351–2359.
[13] L. Li, C. Yang, Y. Zhang, et al., Correctional DP-based energy management strategy of plug-in hybrid electric bus for city-bus route, IEEE Trans. Veh. Technol. 64 (7) (2015) 2792–2803.
[14] G. Li, D. Gorges, Ecological adaptive cruise control and energy management strategy for hybrid electric vehicles based on heuristic dynamic programming, IEEE Trans. Intell. Transport. Syst. 20 (9) (2018) 3526–3535.
[15] C. Sun, F. Sun, H. He, Investigating adaptive-ECMS with velocity forecast ability for hybrid electric vehicles, Appl. Energy 185 (2017) 1644–1653.
[16] Y. Huang, H. Wang, A. Khajepour, et al., Model predictive control power management strategies for HEVs: a review, J. Power Sources 341 (2017) 91–106.
[17] M.J. Gielniak, Z.J. Shen, Power management strategy based on game theory for fuel cell hybrid electric vehicles, IEEE 60th Veh. Technol. Conf. 6 (6) (2004) 4422–4426.
[18] K. Song, F. Li, X. Hu, et al., Multi-mode energy management strategy for fuel cell electric vehicles based on driving pattern identification using learning vector quantization neural network algorithm, J. Power Sources 389 (2018) 230–239.
[19] R. Xiong, Y. Duan, J. Cao, et al., Battery and ultracapacitor in-the-loop approach to validate a real-time power management method for an all-climate electric vehicle, Appl. Energy 217 (2018) 153–165.
[20] Y. Zou, T. Liu, D. Liu, et al., Reinforcement learning-based real-time energy management for a hybrid tracked vehicle, Appl. Energy 171 (2016) 372–382.
[21] T. Liu, B. Wang, C. Yang, Online Markov chain-based energy management for a hybrid tracked vehicle with speedy Q-learning, Energy 160 (2018) 544–555.
[22] X. Hu, T. Liu, X. Qi, et al., Reinforcement learning for hybrid and plug-in hybrid electric vehicle energy management: recent advances and prospects, IEEE Ind. Electron. Mag. 13 (3) (2019) 16–25.
[23] T. Liu, X. Hu, A bi-level control for energy efficiency improvement of a hybrid tracked vehicle, IEEE Trans. Ind. Inf. 14 (4) (2018) 1616–1625.
[24] T. Liu, X. Hu, S.E. Li, et al., Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle, IEEE ASME Trans. Mechatron. 22 (4) (2017) 1497–1507.
[25] Y. Li, H. He, J. Peng, et al., Deep reinforcement learning-based energy management for a series hybrid electric vehicle enabled by history cumulative trip information, IEEE Trans. Veh. Technol. 68 (8) (2019) 7416–7430.
[26] 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, Appl. Energy 211 (2018) 538–548.
[27] J. Yuan, L. Yang, Q. Chen, Intelligent energy management strategy based on hierarchical approximate global optimization for plug-in fuel cell hybrid electric vehicles, Int. J. Hydrogen Energy 43 (16) (2018) 8063–8078.
[28] Y. Wu, H. Tan, J. Peng, et al., Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus, Appl. Energy 247 (2019) 454–466.
[29] J. Wu, H. He, J. Peng, et al., Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus, Appl. Energy 222 (2018) 799–811.
[30] Y. Hu, W. Li, K. Xu, et al., Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning, Appl. Sci. 8 (2) (2018) 187.
[31] X. Qi, Y. Luo, G. Wu, et al., Deep reinforcement learning enabled self-learning control for energy efficient driving, Transport. Res. C Emerg. Technol. 99 (2019) 67–81.
[32] Z. Fu, Z. Li, F. Tao, Adaptive energy management strategy for hybrid batteries/supercapacitors electrical vehicle based on model prediction control, Asian J. Contr. (2019) 1–11.
[33] F. Gao, B. Blunier, A. Miraoui, et al., A multiphysic dynamic 1-d model of a proton-exchange-membrane fuel-cell stack for real-time simulation, IEEE Trans. Ind. Electron. 57 (6) (2010) 1853–1864.
[34] G.J. Osório, E.M.G. Rodrigues, J.M. Lujano-Rojas, et al., New control strategy for the weekly scheduling of insular power systems with a battery energy storage system, Appl. Energy 154 (2015) 459–470.
[35] H. Li, A. Ravey, A. N'Diaye, et al., A novel equivalent consumption minimization strategy for hybrid electric vehicle powered by fuel cell, battery and supercapacitor, J. Power Sources 395 (2018) 262–270.
[36] J. Li, T. Chai, F.L. Lewis, et al., Off-policy interleaved Q-Learning: optimal control for affine nonlinear discrete-time systems, IEEE Trans. Neural Netw. Learn. Syst. (2018) 1–13.

[37] J. Wu, H. He, J. Peng, et al., Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus, Appl. Energy 222 (2018) 799–811.
[38] Z. Fu, Z. Li, P. Si, et al., A hierarchical energy management strategy for fuel cell/battery/supercapacitor hybrid electric vehicles, Int. J. Hydrogen Energy 44 (2019) 22146–22159.
[39] F.A.A. Souza, R. Araújo, J. Mendes, Review of soft sensor methods for regression applications, Chemometr. Intell. Lab. Syst. 152 (2016) 69–79.
[40] L. Di-Bella, S. Fortuna, G. Graziani, et al., A comparative analysis of the influence of methods for outliers detection on the performance of data driven models, IEEE Conf. Instrument. Meas. Technol. (2007) 1–5.
[41] A. Pani, H. Mohanta, A survey of data treatment techniques for soft sensor design, Chem. Prod. Process Model. 6 (1) (2011). Article 2.
[42] J. Han, Y. Park, D. Kum, Optimal adaptation of equivalent factor of equivalent consumption minimization strategy for fuel cell hybrid electric vehicles under active state inequality constraints, J. Power Sources 267 (2014) 491–502.