SCIENCE CHINA
Technological Sciences

•Article•

March 2022 Vol.65 No.3: 713–725
https://doi.org/10.1007/s11431-021-1921-0

Energy management strategy for hybrid electric vehicle integrated with waste heat recovery system based on deep reinforcement learning

WANG Xuan*, WANG Rui, SHU GeQun, TIAN Hua* & ZHANG XuanAng

State Key Laboratory of Engines, Tianjin University, Tianjin 300072, China

Received May 5, 2021; accepted August 16, 2021; published online December 9, 2021

Hybrid electric vehicles (HEVs) are acknowledged to be an effective way to improve the efficiency of internal combustion engines (ICEs) and reduce fuel consumption. Although the ICE in an HEV can maintain high efficiency during driving, its thermal efficiency is approximately 40%, and the rest of the fuel energy is discharged through different kinds of waste heat. Therefore, it is important to recover the engine waste heat. Because of the great waste heat recovery performance of the organic Rankine cycle (ORC), an HEV integrated with an ORC (HEV-ORC) has been proposed. However, the addition of ORC creates a stiff and multi-energy problem, greatly increasing the complexity of the energy management system (EMS). Considering the great potential of deep reinforcement learning (DRL) for solving complex control problems, this work proposes a DRL-based EMS for an HEV-ORC. The simulation results demonstrate that the DRL-based EMS can save 2% more fuel energy than the rule-based EMS because the former provides higher average efficiencies for both engine and motor, as well as more stable ORC power and battery state. Furthermore, the battery always has sufficient capacity to store the ORC power. Consequently, DRL showed great potential for solving complex energy management problems.

hybrid electric vehicles, organic Rankine cycle, waste heat recovery, deep reinforcement learning, energy management system

Citation: Wang X, Wang R, Shu G Q, et al. Energy management strategy for hybrid electric vehicle integrated with waste heat recovery system based on deep reinforcement learning. Sci China Tech Sci, 2022, 65: 713–725, https://doi.org/10.1007/s11431-021-1921-0

1 Introduction

Approximately 20% of the worldwide fuel combustion and associated CO_2 emissions are from transportation [1]. Road transport accounts for approximately 75% of the transportation sector, and the internal combustion engines (ICEs) of heavy-duty vehicles (HDVs) discharge 30% of the on-road CO_2 emissions [2]. Many governments have set reduction targets to reduce the energy consumption and CO_2 emissions of HDVs. For example, the US government projects CO_2 emission reductions of 3%–9% for HDVs by 2027 from the base year 2017 [3]. The Chinese government aims to reduce the fuel consumption of HDVs by 27% from that of a 2012 baseline [4]. Hybrid electric vehicles (HEVs) are acknowledged to be an effective way to improve the efficiency of the ICE and reduce fuel consumption through the use of a hybrid power source [5].

Although the ICE in an HEV can maintain high efficiency during driving, its thermal efficiency is approximately 40% [6]. The remaining fuel energy is discharged through mechanical losses and waste heat from the exhaust and jacket water. Therefore, it is important to recover the engine waste heat. Among various waste heat recovery (WHR) technologies, the organic Rankine cycle (ORC) is regarded as a promising technology because of its great flexibility, low cost, and low maintenance requirements [7,8]. In addition,

*Corresponding authors (email: wangxuanwx@tju.edu.cn; thtju@tju.edu.cn)

previous studies have proven that it has the ability to increase the engine efficiency by approximate 10%–17% [9]. Therefore, an HEV integrated with an ORC WHR system (HEV-ORC) is a promising way to achieve the energy consumption and CO_2 emission reduction targets.

To achieve an HEV with high energy efficiency during actual operation, an energy management system (EMS) for controlling the power distribution of the ICE and motor is critical, and has been a research hotspot for decades. With the addition of the ORC WHR system, the EMS for an HEV-ORC becomes more complex. Biswas and Emadi [10] reviewed the EMS for HEVs (without WHR) based on more than 250 related articles published over the past three decades, and summarized the evolution of the EMS. Initially, an EMS generally utilized a rule-based method [11]. Then, the optimization-algorithm-assisted rule-based method appeared with the goal of finding near optimal rules, followed by global optimal offline control [12]. To implement optimal control in real-time, instantaneous optimal control was proposed with a near-optimal solution [13]. Then, over time, the control methods approached a global optimum solution such as the model predictive control (MPC) method [14]. Recently, deep reinforcement learning (DRL) has been applied to the EMS of HEVs [15], making the EMS closer to global optimal control in real-time for any real-world driving conditions. However, at present, only a few researchers have studied an EMS for HEV-ORC integrated systems.

Feru et al. [16] presented an EMS for an ICE integrated with an electrified ORC WHR system. The EMS consisted of two control levels: the low-level control was for the ORC system, and the high-level was used to determine the settings for parameters such as the distribution of the torques of engine and motor. Both levels utilized MPC. The simulation results demonstrated that comparing with an engine without ORC, the integrated system with the proposed EMS reduced CO_2 emissions by 3.5%. Mansour et al. [17] investigated the fuel saving potential of a mild hybrid electric vehicle coupled with an ORC WHR system for generating electricity, which was stored in a battery. Dynamic programming (DP) was applied for the global optimal EMS. The simulation results indicated that the fuel consumption was reduced by 2.4% on the Worldwide Harmonized Light Vehicles Test Cycles. Kruijt et al. [18] coupled an ORC WHR system with a parallel hybrid HDV and used the heuristic optimal control approach for the EMS. The simulation was carried out under the Urban Dynamometer Driving Schedule and showed a 2.5% improvement in fuel economy.

However, the above EMSs were all based on MPC or DP. The control performance of MPC relies on the precise forecast of future information. Consequently, MPC is not able to guarantee robustness if the actual driving conditions are obviously different from the training conditions [19]. The DP has been used in many EMSs for HEVs, but it cannot be applied in real time [20,21]. As previously mentioned, the introduction of DRL has made the EMS closer to global optimal control in real-time under any real-world driving conditions. DRL combines deep neural networks (DNNs) and reinforcement learning (RL) [22]. Therefore, it has both the strong nonlinear perceptual capability of DNNs and the real-time decision-making ability of RL. Because of these advantages, DRL has shown great potential for solving complex control problems [23], such as the famous Alpha GO [24], which inspired a significant number of DRL applications in practical engineering, including intelligent transportation and EMSs for HEVs without WHR [25–27].

Biswas and Emadi [10] reviewed more than 250 publications on EMSs for HEVs from 1993 to 2018 and pointed out that the RL is an important development trend of EMS in the future. Zou et al. [28] put forward an online-updated EMS based on deep Q learning and validated through hardware-in-the-loop simulation. Lian et al. [29] proposed an improved energy management framework that embeds expert knowledge into DRL, the proposed framework not only accelerates the learning process, but also gets a better fuel economy. However, there are still some problems in the application of DRL in EMS, such as poor robustness and stability. In the future, the improvement of algorithm, the progress of hardware facilities and the progress of cloud computing technology will be the important means to realize the application of DRL.

To the best of the author's knowledge, DRL has not been applied to the EMS of an HEV-ORC. In contrast to an HEV without the ORC, in which both the ICE and motor can respond quickly, the ORC responds much more slowly than the ICE and motor [30], making it a stiff problem. In addition, when the ORC is incorporated, the whole system contains more kinds of energy, including heat (ORC and ICE), electricity (generator and ORC), mechanical (ICE and motor), and chemical energy (ICE and battery). Therefore, the addition of the ORC creates a stiff and multi-energy problem, and greatly increases the complexity of the EMS. Considering the great potential of DRL for solving complex control problems, as previously mentioned, this study first applies DRL to the EMS of an HEV-ORC to prove its ability for solving stiff and multi-energy problems. To reveal the advantages of the DRL-based EMS, it is compared with the traditional rule-based EMS under highway conditions. It is believed that this work is a good reference for applying DRL to solve other complex, stiff, and multi-energy problems. The rest of the paper is organized as following.

Section 2 briefly describes the structure of the HEV-ORC and the DRL-based energy management, including the rule-based EMS and operation strategy of the ORC. Section 3 presents the detailed mathematical models of the HEV-ORC integrated system and the DRL-based EMS. The simulation results for the DRL-based EMS and rule-based EMS are

compared and discussed in Section 4. Finally, the conclusions are presented in Section 5.

2 Description of structure of HEV-ORC and EMS

2.1 Structure of HEV-ORC

In this work, the HEV-ORC contains a non-plug-in HEV with a single-shaft parallel hybrid powertrain structure and an ORC WHR system, as shown in the environment part of Figure 1. The ICE and motor have the same main axis and the same speed, while they can operate independently or together. The motor consumes electricity from the battery to drive the vehicle, or works as a generator driven by the ICE to generate electricity, which is stored in the battery. This hybrid powertrain structure is potential for practical applications, owing to the simple layout and high degree of freedom [31]. The ORC recovers the waste heat of the exhaust and outputs electricity, which is stored in the battery.

The basic principle of the ORC is as follows: the heat source (exhaust) heats the working fluid into a gas with a high temperature and pressure in the evaporator. This gaseous working fluid then expands in the turbine of a generator and generates electrical power. Subsequently, the expanded gas is condensed back into a liquid and pumped to the evaporator again, restarting the next cycle. Toluene is selected as the working fluid because of its high efficiency, environmental friendliness, and low cooling source requirement [32]. The jacket water is adopted as the cooling source of the ORC, which cools the engine first and then the working fluid, and finally is cooled by the air in the vehicle air-radiator. It is necessary to maintain a small degree of superheat for the dry working fluid at the turbine inlet during the entire operation to protect the turbine blades [33]. The mass flow rate of the working fluid is typically controlled by the pump speed to track the reference degree of superheat [34]. Therefore, a proportional-integral-derivative (PID) controller is used to control the pump speed and maintain the small degree of superheat, as shown in Figure 1. The ORC operation strategy is as follows.

(1) At the beginning of the ORC, the turbine is bypassed by the bypass valve. After the working fluid absorbs enough waste heat and the superheat becomes greater than zero, the turbine is connected to generate electricity.

(2) When the engine exhaust heat is below the lower limit or the PID controller fails to maintain the superheat above zero, the turbine is bypassed to ensure safety. At the same

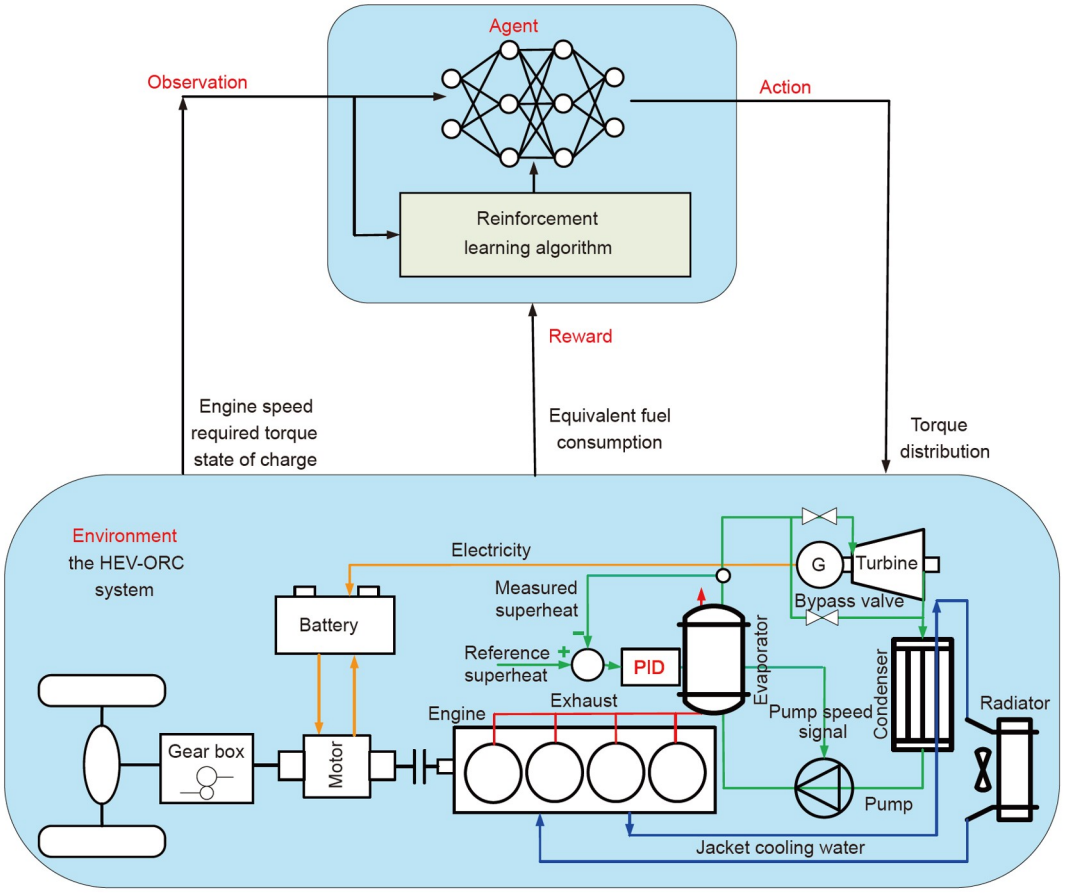

Figure 1 (Color online) Structure of HEV-ORC integrated system and basic schematic of DRL-based EMS.

time, the ORC still works but without output power, and the superheat of the working fluid at the inlet of the bypass valve is controlled by the PID controller to approach the reference superheat.

(3) When the engine exhaust heat has enough heat and the superheat is controlled by the PID controller above a certain value, the turbine is again connected to produce electricity.

(4) When the engine stops, the ORC stops as well.

The main parameters of the HEV-ORC are cited from our previous work [35] and listed in Table 1.

2.2 Fundamentals of DRL-based EMS

DRL combines a DNN and RL. As a goal-directed algorithm, the DRL agent learns to perform a task well through interacting with the environment. The agent consists of two components: the policy and learning algorithm. The policy is a mapping between the optimal actions and observations from the environment, which is a DNN with tunable parameters in DRL. The learning algorithm continuously updates the parameters of the DNN policy based on the observations, actions, and rewards generated from repeated training episodes. Finally, through repeated training, an optimal policy is acquired for taking a series of actions to maximize the cumulative reward, without being explicitly programmed and without human intervention during the task. Figure 1 presents the basic scheme of the DRL-based energy management for an HEV-ORC.

As shown in Figure 1, the objective of the DRL-based EMS is to obtain the minimum energy consumption. The agent receives observations (engine speed, required torque) and rewards (less energy consumption corresponds to a larger reward) from the environment (the dynamic model of the HEV-ORC) and sends actions (torque distribution) to the environment. Through repeated training, the agent learns the optimal strategy for torque distribution. Each part of the DRL-based EMS shown in Figure 1 is described in detail in Section 3, which outlines the mathematical model.

2.3 Fundamentals of rule-based EMS

The rule-based EMS has the goal of maintaining the engine efficiency in an optimal region and is based on our previous research [35]. The vehicle operation process is divided into eight working modes: 0-parking, 1-brake energy recovery, 2-mechanical brake without brake energy recovery, 3-motor drive, 4-engine drive, 5-hybrid drive, 6-driving charge, and 7-forced discharge. Table 2 lists the detailed rules of this strategy, where T indicates the torque, and SOC is the state of charge in the battery. The subscripts tar, e, m, and opt denote the target, engine, motor, and optimal values, respectively. T_m can be positive or negative, which means working as a generator or motor. It should be noted that the forced discharge mode is used to maintain sufficient space all along to store the recovered power generated by the ORC. When entering the forced discharge mode, if the SOC falls below a set value (SOCdis), the working mode ends. Table 3 lists the limits and calculation methods for each boundary parameter. More details about the rule-based EMS can be found in our previous work [35].

3 Mathematical model

The dynamic model of the HEV-ORC is the environment and established using Matlab-Simulink. The DNN policy and

Table 1 Main parameters of HEV-ORC [35]

Component	Parameter	Value
Vehicle	Frontal area (m^2)	6.45
	Vehicle weight (kg)	8100
	Wheel radius (m)	0.5
Battery	Type	Lithium ion
	Capacity (kW h)	3.3
Motor	Maximum power (kW)	88
	Maximum torque (N m)	200
Engine	Type	Diesel engine
	Number of cylinders	4
	Rated power (kW)	169
	Rated speed (r/min)	2200

Table 2 Control rules for hybrid powertrain control strategy

Logical threshold		Mode	Torque distribution
$T_{tar} < 0$	$SOC < SOC_{max}$	1	$T_m = -T_{tar}$, $T_e = 0$
	$SOC \geq SOC_{max}$	2	$T_e = T_m = 0$
$T_{tar}=0$		0	$T_e = T_m = 0$
$0 < T_{tar} < T_{emin}$	$SOC > SOC_{min}$	3	$T_e = 0$, $T_m = -T_{tar}$
	$SOC \leq SOC_{min}$	4	$T_e = T_{tar}$, $T_m = 0$
$T_{emin} \leq T_{tar} < T_{eopt}$	$SOC < SOC_{max}$	6	$T_e = T_{eopt}$, $T_m = T_{eopt}-T_{tar}$
	$SOC = SOC_{max}$	7	$T_e = 0.8T_{tar}$, $T_m = -0.2T_{tar}$
	$SOC \geq SOC_{max}$	4	$T_e = T_{tar}$, $T_m = 0$
$T_{eopt} \leq T_{tar}$	$SOC > SOC_{min}$	5	$T_e = T_{eopt}$, $T_m = T_{eopt}-T_{tar}$
	$SOC \leq SOC_{min}$	4	$T_e = T_{tar}$, $T_m = 0$

Table 3 Boundary condition for rule-based EMS

Parameter	Value or calculation method
SOC_{max}	0.9
SOC_{min}	0.3
SOC_{dis}	0.8
T_{eopt}	A function of engine speed
T_{emin}	A function of engine speed

learning algorithm are also based on Matlab-Simulink. The entire dynamic model of HEV-ORC is composed of numerous subsystems such as the ICE, motor, battery, and ORC, as shown in Figure 2.

3.1 HEV-ORC model (environment)

(1) HEV model

The HEV model has been used in many studies [25–27], and the model in this work is the same as that used in our previous research [35]. Therefore, only a brief introduction of each part of the HEV model is provided here. More details about the main equations and data can be found in ref. [35].

Based on the experimental data of the object ICE, the engine fuel consumption, efficiency, exhaust temperature, and exhaust flow rate are fitted as functions of the engine speed and torque. A first-order inertia element is added in the function to represent the dynamic characteristic.

The motor works in parallel with the ICE. It acts as a motor to drive the vehicle together with ICE or by itself, and also works as a generator to generate electricity, which is stored in the battery. Because of its fast dynamic response speed, a static model is adopted.

The model of battery pack is typically established as a circuit for simplification, and composed of a dynamic capacitor voltage, internal resistance voltage, and open circuit voltage.

Similar to the basic principle of a proportional-integral (PI) controller, the driver presses the acceleration pedal or brake pedal based on an observation of a deviation from the objective speed. Therefore, the driver model is typically established as a PI controller.

The vehicle dynamics model is used to convert the torques of the ICE and motor into driving forces to overcome the rolling resistance, air resistance, gravity resistance, and acceleration resistance. According to the resultant of the driving force and resistances, the speed of the vehicle is calculated.

(2) ORC model

Because the ORC WHR system is regarded as more suitable for highway conditions [36], as a result of the relatively stable and large engine load, the simulation of the HEV-ORC system is conducted under a highway driving cycle. The ORC is designed for a medium load on the ICE, which is the most common operating condition for the ICE when driving on highway conditions. The exhaust temperature and flow rate under the design condition are 450°C and 0.15 kg/s, respectively. As previously mentioned, the cooling source of the ORC is the jacket water of approximately 90°C. Based on these boundary conditions, the evaporating pressure and condensing temperature are designed to be 2 MPa and

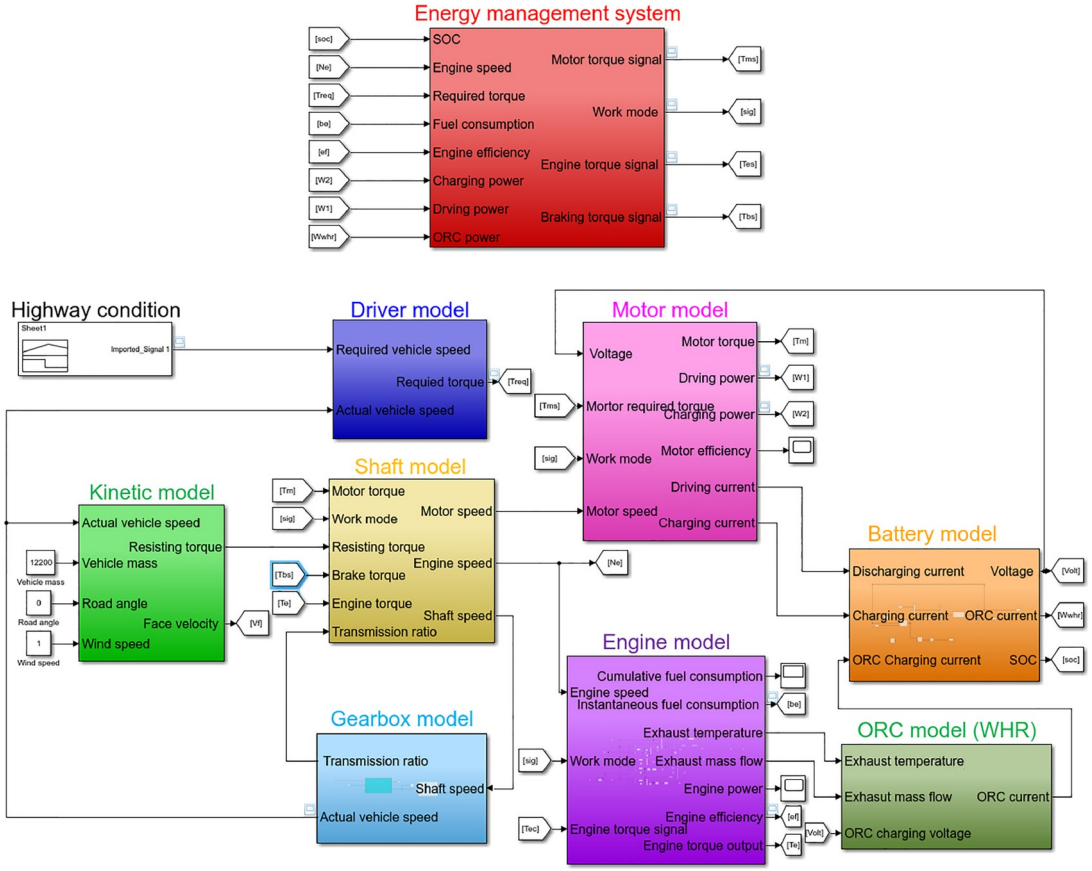

Figure 2 (Color online) Dynamic model of HEV-ORC in simulink.

120°C, respectively, with the consideration of a reasonable pressure ratio for such a small turbine.

The ORC consists of four main components (evaporator, condenser, turbine, and pump), as shown in Figure 1. The models of the main components are first established and then connected together based on their interrelationships to create the entire dynamic model of the ORC. The detailed modeling process and validation were described in detail in our previous research [37]; therefore, only a brief introduction is given here.

The models of the evaporator and condenser are established using the moving boundary (MB) method. Figure 3 shows the MB model of the evaporator as an example. The evaporator is simplified as a convection heat transfer device, and divided into three regions (sub-cooling, two-phase, and super-heating regions) because of the significantly different heat transfer coefficients in these three phase regions. During the simulation, the length of each region is tracked all along. The lumped parameter method is used in each region, and the general energy and mass balance equations can be derived as eqs. (1)–(3).

General mass balance equation for the three regions:

$$\int_0^{L_i} \frac{\partial(A\rho)}{\partial t}\mathrm{d}z + \int_0^{L_i} \frac{\partial m}{\partial z}\mathrm{d}z = 0. \tag{1}$$

General energy balance equation for the three regions:

$$\int_0^{L_i} \frac{\partial(A\rho h - Ap)}{\partial t}\mathrm{d}z + \int_0^{L_i} \frac{\partial mh}{\partial z}\mathrm{d}z = \int_0^{L_i} \alpha_i \pi D_i (T_w - T_r)\mathrm{d}z. \tag{2}$$

Simplified energy balance equation of the tube wall:

$$c_{pw}\rho_w A_w \frac{\mathrm{d}T_w}{\mathrm{d}t} = \alpha_i \pi D_i (T_r - T_w) + \alpha_o \pi D_o (T_a - T_w). \tag{3}$$

The MB model for the working fluid is developed by integrating eqs. (1)–(3) along the length of each region. Corresponding to the three regions of the working fluid, the heat source is divided into three regions with the same length, and its MB model is established in the same way.

Because the pump and turbine respond much faster than heat exchangers, their models are usually represented by static models in the system model for simplification. A displacement pump is used, and the mass flow rate is calculated using eq. (4) [38]. η_v, V_{cyl}, ω, and ρ_{pump} are the volumetric efficiency, cylinder volume, pump speed, and density of the working fluid, respectively.

$$m_{pump} = \eta_v \cdot V_{cyl} \cdot \omega \cdot \rho_{pump}. \tag{4}$$

The mass flow rate model for a turbine is typically simplified as a nozzle [39], as described in eq. (5). C_v is a coefficient. ρ_{in} and p are the working fluid density at the turbine inlet and the evaporating pressure, respectively.

$$m_t = C_v \sqrt{\rho_{in} p}. \tag{5}$$

The calculations of the working fluid enthalpy at the outlet of the pump and turbine are similar, as shown in eqs. (6) and (7), respectively, where the subscripts p and t denote the pump and turbine, respectively. The subscript s indicates an isentropic process such as the isentropic enthalpy or efficiency.

$$h_{pout} = h_{pin} + (h_{spout} - h_{pin}) / \eta_{sp}, \tag{6}$$

$$h_{tout} = h_{tin} - (h_{tin} - h_{stout})\eta_{st}. \tag{7}$$

The dynamic model of the ORC system is established by connecting the above component models together.

(3) Simplification of ORC model

The original ORC model requires a relatively large amount of computational resources because of the nonlinearity of the heat exchanger model. During the DRL training process, the dynamic model of the entire HEV-ORC needs to be calculated many times. To save training time, the ORC model should be simplified when training. The original model is linearized at the design point using the exhaust temperature, exhaust flow rate, and pump speed as inputs, and the net output power and degree of superheat as outputs. The degree of superheat output is used to detect whether the superheat is controlled above 0°C because if there is no superheat, the turbine needs to be bypassed to protect the turbine blades.

It is well known that a linearized model is close to the original model around the linearized point, but the difference between the linearized model and the original model becomes increasingly large as the operating point moves away from the linearized point. Figure 4(a) compares the output power of the linearized model and the original model under a large variation in the engine exhaust. The corresponding variations in the exhaust temperature and mass flow rate are shown in Figure 4(b). It can be observed that under most conditions, especially around the design condition, the linearized model maintains an acceptable accuracy. Because this work considers highway conditions, the output power of the ICE is relatively stable, as well as the exhaust parameters and ORC output power, as shown below. Consequently, the linearized model is reasonable for highway conditions.

3.2 DRL-based EMS

(1) Action, observation, and reward

The torque distribution is the core of the EMS. Therefore, the torque of the ICE is selected as the action, and the torque of the motor is calculated using the difference between the ICE torque and the total demand torque for the driving vehicle, which is obtained using the HEV model. To save braking energy, when the vehicle is braking, the motor works as a generator to match the brake torque demand and generate electricity, which is stored in the battery, except when the SOC of the battery is below the lower limit.

The observations should represent the status of the HEV-ORC. According to the observed status, the DRL agent can

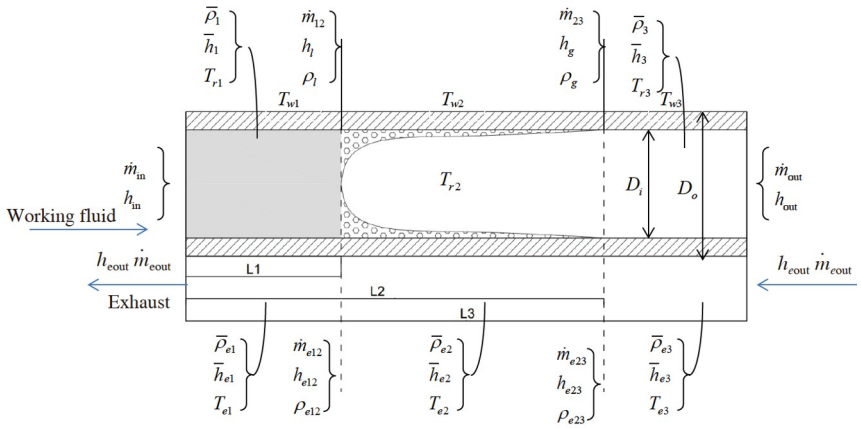

Figure 3 (Color online) Schematic of MB model of evaporator [37].

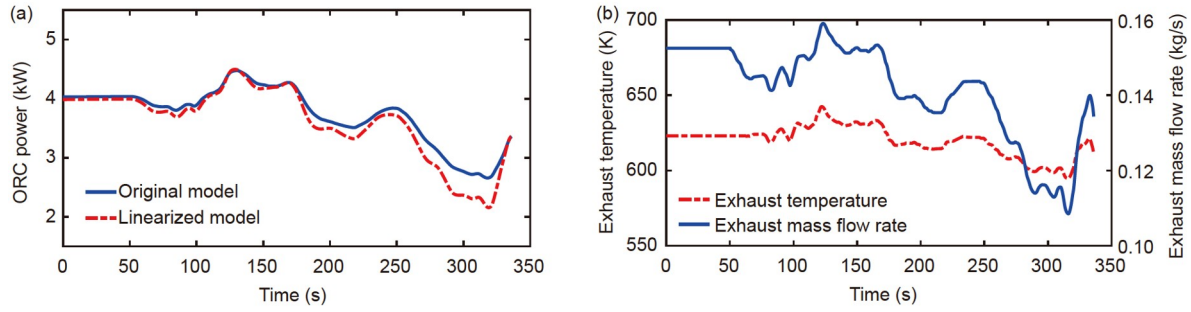

Figure 4 (Color online) Differences between original model and linearized model. (a) ORC power comparison between original and linearized models under large variation in engine exhaust; (b) corresponding variations in exhaust temperature and mass flow.

make the optimal decision. The vehicle speed, vehicle power demand, vehicle torque demand, and SOC of the battery are common observations in the literature on DRL-based EMSs for HEVs without ORCs [15]. The SOC is used to determine the status of the battery. The vehicle power demand is calculated using the vehicle speed and torque demand. Therefore, two of the three variables are independent. Because the engine and motor have the same speed, which is related to the vehicle speed, the given SOC, vehicle speed, and torque demand can be used to determine the optimal status of the engine and motor. The status of the ORC is related to the status of the engine exhaust, which is determined by the engine torque and speed. Therefore, the vehicle speed, vehicle torque demand, and SOC of the battery are the observations used to represent the status of the HEV-ORC in this study.

Immediate rewards are critical for a DRL-based EMS. The RL algorithm has the goal of obtaining the maximum cumulative reward using a series of the best actions. Therefore, the reward should be defined in accordance with the optimization objective [40]. In this study, the designed reward is obtaining the minimum total energy consumption for the ICE and motor, which is called the equivalent fuel consumption [41]. The equivalent fuel consumption is calculated using eq. (8). Here, S represents the equivalent factor of the charge or discharge [15].

$$C_{\mathrm{eq}} = P_{\mathrm{ICE}}/\eta_{\mathrm{ICE}} - P_{\mathrm{ORC}}/\eta_{\mathrm{ICE}} + P_{\mathrm{discharge}}/S_{\mathrm{discharge}} - P_{\mathrm{charge}}/S_{\mathrm{charge}}. \quad (8)$$

To maintain the SOC of the battery within the upper and lower limits, when the SOC exceeds the boundary during training, the training episode stops immediately and starts the next episode with the return of a large punishment. In addition, it is well known that frequently starting and stopping harms an ICE and a motor. Thus, there is a punishment item for start/stop events in the reward function. Finally, the reward function is shown in eq. (9). The addition of a constant, c, in eq. (9) ensures that the reward is always a positive value [15]. However, based on the research experience in this study, the value of c is critical and sensitive to the convergence of the DRL training process. Its value should not be much greater than the sum of the last three items in eq. (9). Otherwise, the training process may not converge.

$$R = c - C_{\mathrm{eq}} - 100(\mathrm{SOC} > 0.9 \,||\, \mathrm{SOC} < 0.3) - 10(\mathrm{sig}_t \neq \mathrm{sig}_{t+1}). \quad (9)$$

(2) Agent

RL algorithms include policy- and value-based methods. A

deep Q-value network (DQN) is a typical value-based algorithm. A DQN uses a network to approximate the mapping between the Q values (long-term reward) and the state-action pairs. According to this mapping, the agent takes the action with the maximum Q value for each step. In this work, the DQN algorithm is adopted. The action of the engine torque is discrete and the discretization resolution is 1 N m. To improve the stability of the DQN, two networks are used. Critic Q (S, A) takes observations and actions as inputs and outputs the corresponding Q value. Target critic Q' (S, A) is used to improve the stability of the DQN, and its parameters are periodically updated according to the latest parameters of the critic. Both networks have the same structure and parameters. The final trained Q-value approximator is stored in the critic network. The algorithm framework of the DQN is shown in Figure 5. During the training process, the parameters of the two networks are initialized with the same value, and then the agent repeats the following steps until it converges.

(1) For current observation S, randomly select an action with probability ε. Otherwise, select the action with the greatest Q-value.

(2) Execute action A. Observe reward R and next observation S'.

(3) Store the experience (S, A, R, S') in the replay memory buffer.

(4) Randomly, sample M experiences for a mini-batch of (S_i, A_i, R_i, S'_i) from the experience buffer.

(5) Update the Q-value with the two networks in eqs. (10) and (11).

$$A_{\max} = \operatorname{argmax}_A Q(S_i', A' \mid \theta^Q), \tag{10}$$

$$y_i = R_i + \gamma Q'(S_i', A_{\max} \mid \theta^{Q'}). \tag{11}$$

(6) Update the network parameters of the critic by one-step minimization of the loss function of eq. (12) across all samples from the mini-batch.

$$L = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(S_i, A_i \mid \theta^Q)^2. \tag{12}$$

(7) Update the network parameters of the target according to the latest parameters of the critic.

The neural-network structure of the Q network is shown in Figure 6. It is established using the MATLAB deep learning toolbox. Each fully connected layer has 48 neurons. The number of neurons and the structure of the network greatly affect the training performance, and they depend on the complexity of the specific problem. The main parameters for creating and training the DQN agent are listed in Table 4. The training performance is sensitive to these parameters. For example, different values for the discount factor produce different training results, some of which cannot converge. Discount factors between 0.9 and 0.99 are tested in this work, and it is found that 0.99 yields the largest cumulative reward.

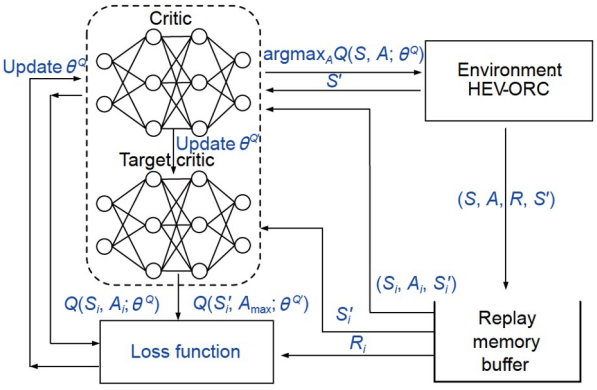

Figure 5 (Color online) Algorithm framework of DQN.

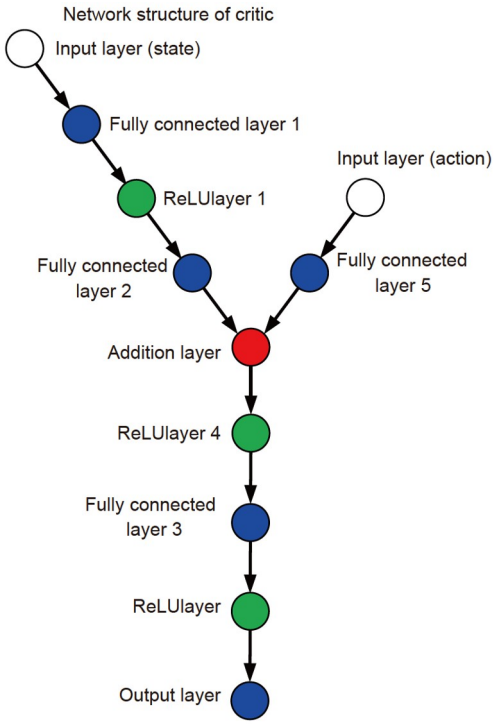

Figure 6 (Color online) Structure of Q network.

Table 4 Main parameters for creating and training DQN agents

Parameter	Value
Discount factor	0.99
Greedy exploration ε	1
Learning rate	10^{-3}
Gradient threshold	1
Sample time	5
Target smooth factor	10^{-3}
Mini-batch size	64
Experience buffer length	10^6

4 Results and discussion

This section compares the simulation results for the DRL-based EMS and rule-based EMS under the standard highway driving cycle to prove the potential of DRL for the EMS of the HEV-ORC system. The highway driving cycle is selected because the ORC WHR system is regarded as more suitable for highway conditions.

The target vehicle speed and simulated vehicle speed are compared in Figure 7, which demonstrates good consistency and verifies the rationality of the parameters used in this HEV-ORC system model. Figure 8 shows the training process for the DRL-based EMS. Before training, the agent has no experience regarding the distribution of the torque. Therefore, the agent requires numerous attempts to learn the EMS. To improve the convergence of the training process, some studies have used a pre-trained network with optimal experience samples before training [42]. However, because of the extensive exploration without human intervention, the DRL agent may obtain a larger cumulative reward [43].

Figure 9 shows a comparison of the equivalent fuel consumption values of the DRL-based EMS and rule-based EMS. It can be observed from Figure 9(a) that the equivalent fuel consumption with the DRL-based EMS is not always lower than that with the rule-based EMS and sometimes even slightly higher. This is because the optimization objective is to obtain the largest cumulative reward. RL is a powerful family of DP, and it makes the decision that optimizes the system performance over the entire process. Figure 9(b) shows the cumulative equivalent fuel consumption values with the two EMSs. At last, the DRL-based EMS and rule-based EMS consume 125.2 and 122.7 MJ, respectively. The DRL-based EMS saves 2% of the fuel energy compared to the rule-based EMS. The fuel consumption is determined by the engine efficiency, motor efficiency, and ORC performance. Therefore, these parameters are demonstrated and analyzed in the text below.

Figure 10 shows the variation in the engine torques under the test highway conditions. The engine torque with the DRL-based EMS is relatively stable, while that with the rule-based EMS is volatile. The rule-based EMS has the goal of maintaining the optimal engine efficiency, but the optimal engine torque is usually larger than the required torque. To reach the optimal engine torque, the vehicle usually needed to work in the driving charge mode, in which the abundant engine torque is used to charge the battery. This means that the SOC of the battery is often close to the upper limit, and the energy recovered by the ORC cannot be used to charge

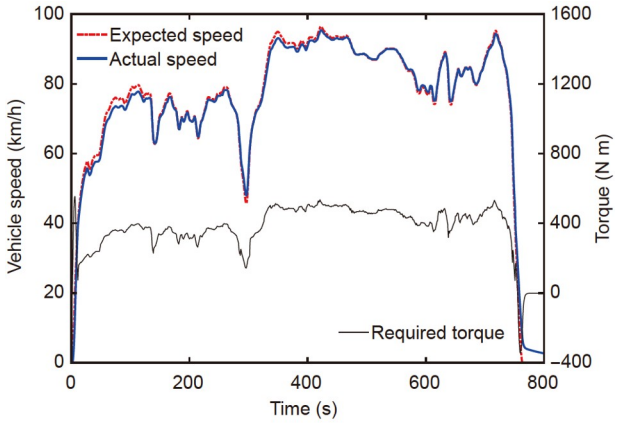

Figure 7 (Color online) Target vehicle speed and simulated vehicle speed.

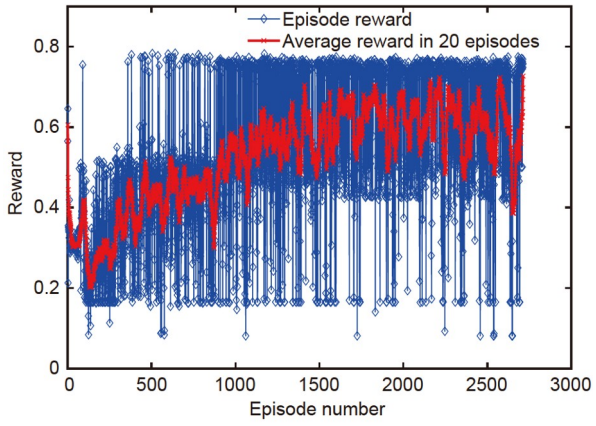

Figure 8 (Color online) Training process of DRL-based EMS of HEV-ORC system.

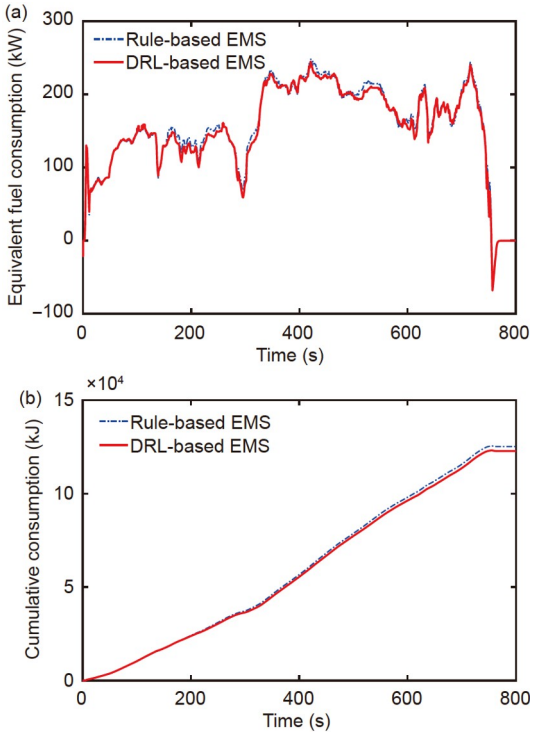

Figure 9 (Color online) Fuel consumption comparison between DRL-based and rule-based EMS. (a) Fuel consumption at every step; (b) cumulative fuel consumption.

the battery, as shown in Figure 11. To maintain sufficient space in the battery all along for storing the power recovered by the ORC, the vehicle often needs to work in the forced discharge mode. Therefore, the engine torque with the rule-based EMS is volatile, as well as the engine efficiency, as shown in Figure 12.

In contrast, the engine torque with the DRL-based EMS is usually less than the optimal torque, which prevents its engine efficiency from reaching the optimal value. However, the engine efficiency remains relatively high and stable. Although the engine efficiency with the DRL-based EMS is sometimes less than that with the rule-based EMS, as shown in Figure 12, the average engine efficiency with the former (0.41) is larger than that with the latter (0.40). As previously mentioned, the DRL-based EMS makes decisions that consume the least amount of cumulative fuel energy over the entire process. As shown in Figure 11, the SOC variation with the DRL-based EMS stays at approximately 0.65 at most times, which is much more stable than that with the rule-based EMS and more beneficial for the battery. Because the SOC is always evidently lower than the upper limit, it does not need to work in the forced discharge mode, in which the engine efficiency is relatively low.

Figure 13 shows the motor torque variations with the DRL-based EMS and rule-based EMS. Both of them change drastically, but the off-design efficiency of the motor is more stable and higher than that of the engine, as shown in Figure 14. The average motor efficiency of the rule-based EMS is 91.85%, while that of the DRL-based EMS is 92.89%, which is slightly higher than the former. In summary, the average efficiencies of both engine and motor are improved with the DRL-based EMS compared to the rule-based EMS. The DRL-based EMS is long-term optimization, while the rule-based EMS is instantaneous optimization. Thus, even though the efficiencies of the engine and motor with the DRL-based EMS are not always higher, the average efficiencies are higher.

Figure 15 shows the ORC output power values with the two EMSs. Similar to the engine torque, the ORC power of the DRL-based EMS is much more stable than that of the rule-based EMS because the ORC power is positively related to the exhaust waste heat, which is also positively related to the engine torque. Sometimes, the ORC power of the rule-based EMS is zero because the superheat of the ORC is below zero, as shown in Figure 16. When the engine torque changes sharply, the exhaust waste heat varies too quickly so

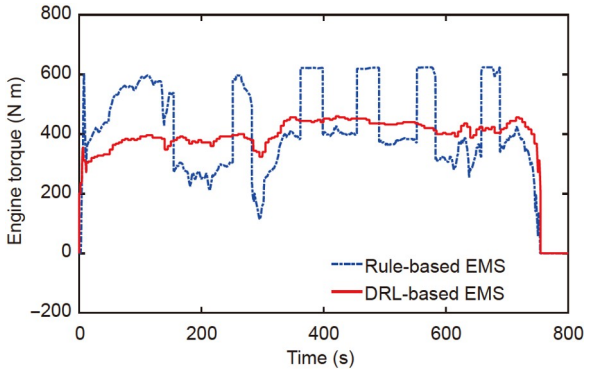

Figure 10 (Color online) Engine torques with DRL-based EMS and rule-based EMS.

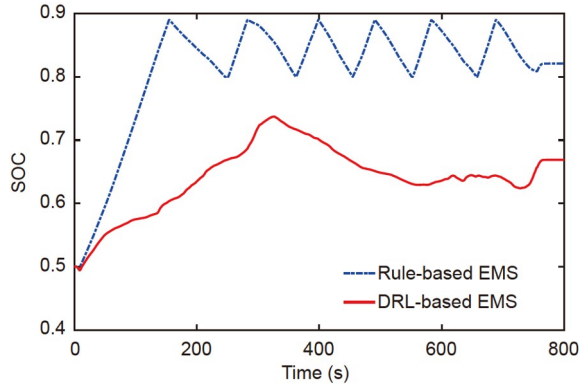

Figure 11 (Color online) SOC variations with DRL-based EMS and rule-based EMS.

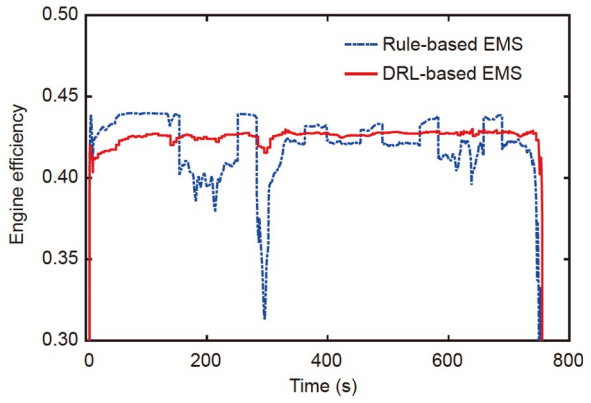

Figure 12 (Color online) Engine efficiencies with DRL-based EMS and rule-based EMS.

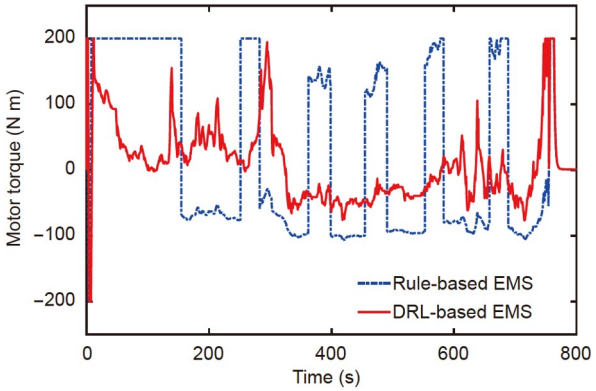

Figure 13 (Color online) Motor torques with DRL-based EMS and rule-based EMS.

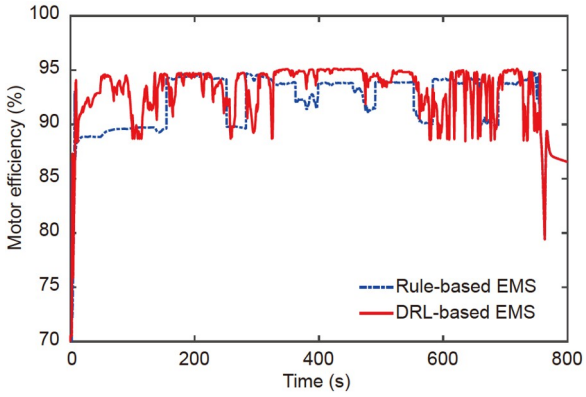

Figure 14 (Color online) Motor efficiencies with DRL-based EMS and rule-based EMS.

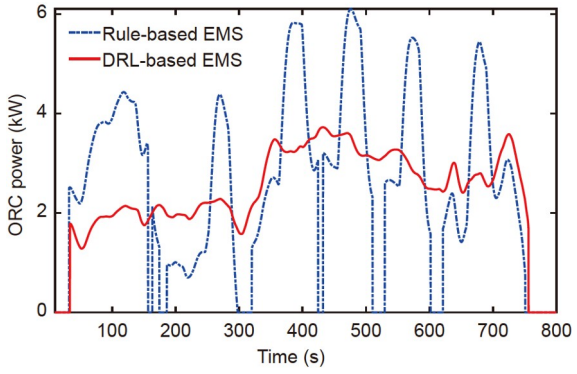

Figure 15 (Color online) ORC outputs with DRL-based EMS and rule-based EMS.

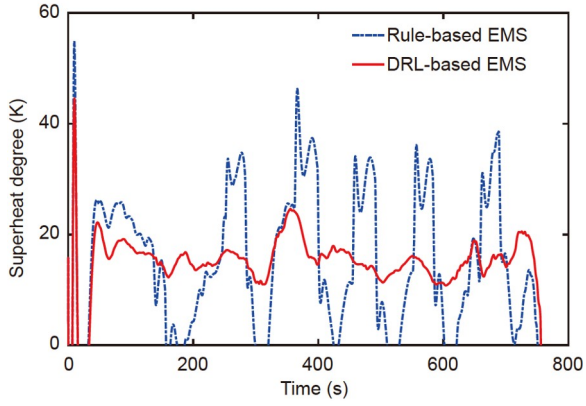

Figure 16 (Color online) Degrees of superheat for ORC with DRL-based EMS and rule-based EMS.

that the PID controller cannot immediately establish an ORC superheat value greater than zero. Therefore, to protect the turbine blades, the expander is bypassed and the ORC power becomes zero. When the PID controller is able to again establish a superheat value greater than zero, the turbine is connected to produce power.

The ORC power is continually stored in the battery. Therefore, the battery needs to maintain sufficient space all the time for storing the power recovered by the ORC to improve the entire energy utilization rate. As previously mentioned, an important reason for the violent fluctuations in engine torque is the use of the forced discharge mode to free battery space. In brief, the forced discharged mode assists in maintaining sufficient battery capacity, but it does harm to the power recovered by the ORC and the engine efficiency. In the HEV-ORC system, the critical problem is to deal with the complex relationships among the engine, motor, battery, and ORC, to find an optimal operation strategy that can maintain the high efficiency of each component and sufficient battery capacity to store the recovered power all along. Based on the previous analysis, the rule-based EMS does not provide a satisfactory tradeoff among these components.

Because these components are closely related and involve different forms of energy with obviously different characteristics, it is difficult to find the optimal operation strategy. Specifically, because the engine power and motor power respond in several seconds, they are often modeled as having instantaneous responses. In contrast, the ORC requires a much longer response time (several minutes). Because of the great differences in the dynamic response speeds of the engine, motor, and ORC, the model of the HEV-ORC becomes a stiff problem, and the optimal strategy should consider a long-term plan. In addition, there should be sufficient space in the battery all along to store the recovered power. Thus, the optimization of the battery operation is more complex than that for a normal HEV. Furthermore, the distribution of the engine torque and motor torque should also be optimized just as with the EMS of a normal HEV. Human experience is powerless for such a complex, stiff, and multi-energy problem. Thus, a rule-based EMS does not perform very well. In particular, before simulating the HEV-ORC with the rule-based EMS, it may not be predicted that such a fluctuating engine torque would produce exhaust waste heat that could not always be recovered.

In contrast, the DRL-based EMS can detect this problem through repeated training episodes. As shown in Figures 9(a), 15, and 16, the engine torque remains relatively stable and the ORC superheat can be controlled above zero by the PID controller all along. Thus, the ORC outputs power continually by recovering exhaust waste heat. In addition, the DRL-based EMS can ensure that the battery always has enough space to store the recovered power, and the variation in the battery SOC is much more stable than that with the rule-based EMS. At the same time, the DRL-based EMS provides a better distribution of the engine torque and motor torque because the average efficiencies of both the engine and motor are higher than those with the rule-based EMS. In summary, without any human experience, the DRL agent learns a satisfactory strategy for such a complex, stiff, and multi-energy problem and provides more comprehensive consideration than the rule-based EMS. The satisfactory

performance of the DRL-based EMS demonstrates the great potential of DRL for solving complex energy management problems.

5 Conclusions

This study establishes a dynamic simulation model of an HEV-ORC integrated system. Because of the addition of the ORC to the HEV, the integrated system contains many kinds of energy, and the components have evidently different dynamic response characteristics. A DRL-based EMS is proposed for such a complex, stiff, and multi-energy problem because of the great potential of DRL for solving complex control problems. To reveal the advantages of the DRL-based EMS, it is compared with a rule-based EMS under highway conditions.

The simulation results demonstrate that the DRL-based EMS and rule-based EMS consume 125.2 and 122.7 MJ of cumulative equivalent energy, respectively. The DRL-based EMS can save 2% of the fuel energy compared with the rule-based EMS. The average efficiencies of both the engine and motor are higher than those with the rule-based EMS. In addition, the variations in the engine torque and battery SOC with the DRL-based EMS are more stable than those with the rule-based EMS. Therefore, the heat source of the ORC (engine exhaust) is also stable, and the ORC can safely be controlled to continually generate power by recovering waste heat. At the same time, there is always sufficient space in the battery for storing the ORC power. In contrast, the rule-based EMS often needs to operate in the forced discharge mode to free space in the battery for storing the ORC power. Thus, the engine torque fluctuates violently, and it is often necessary to cut off the ORC power to protect the turbine blades.

In summary, through repeated training, the DRL agent can more comprehensively consider the tight interrelationships among the components in the HEV-ORC integrated system without any human experience, compared to the rule-based EMS. The satisfactory performance of the DRL-based EMS shows the great potential of DRL for solving complex, stiff, and multi-energy problems. Therefore, it is believed that DRL will play an important role in the future for the EMSs of extremely complex and large energy systems.

This work was supported by the National Natural Science Foundation of China (Grant No. 51906173).

1 The World Bank. CO_2 emissions from transport (% of total fuel combustion). https://data.worldbank.org/indicator/EN.CO_2.TRAN.ZS
2 EIA. International Energy Outlook. 2016. https://www.eia.gov/outlooks/aeo/er/
3 Di Battista D, Fatigati F, Carapellucci R, et al. Inverted Brayton Cycle for waste heat recovery in reciprocating internal combustion engines. Appl Energy, 2019, 253: 113565
4 The ICCCT. CO_2 Emission Standards for Passenger Cars and Light-Commercial Vehicles in the European Union. Technical Report. January 2019
5 Faraj M, Basir O. Range anxiety reduction in battery-powered vehicles. In: 2016 IEEE Transportation Electrification Conference and Expo (ITEC). Dearborn: IEEE, 2016
6 Abedin M J, Masjuki H H, Kalam M A, et al. Energy balance of internal combustion engines using alternative fuels. Renew Sustain Energy Rev, 2013, 26: 20–33
7 Yu G, Shu G, Tian H, et al. Multi-approach evaluations of a cascade-Organic Rankine Cycle (C-ORC) system driven by diesel engine waste heat: Part B-techno-economic evaluations. Energy Convers Manage, 2016, 108: 596–608
8 Algieri A, Morrone P. Comparative energetic analysis of high-temperature subcritical and transcritical Organic Rankine Cycle (ORC). A biomass application in the Sibari district. Appl Thermal Eng, 2012, 36: 236–244
9 Shu G, Wang X, Tian H. Theoretical analysis and comparison of Rankine cycle and different organic Rankine cycles as waste heat recovery system for a large gaseous fuel internal combustion engine. Appl Thermal Eng, 2016, 108: 525–537
10 Biswas A, Emadi A. Energy management systems for electrified powertrains: State-of-the-art review and future trends. IEEE Trans Veh Technol, 2019, 68: 6453–6467
11 Lee H D, Sul S K. Fuzzy-logic-based torque control strategy for parallel-type hybrid electric vehicle. IEEE Trans Ind Electron, 1998, 45: 625–632
12 Piccolo A, Ippolito L, Galdi V, et al. Optimisation of energy flow management in hybrid electric vehicles via genetic algorithms. In: ASME International Conference on Advanced Intelligent Mechatronics. Como: IEEE, 2001
13 Pérez L V, Bossio G R, Moitre D, et al. Optimization of power management in an hybrid electric vehicle using dynamic programming. Math Comput Simul, 2006, 73: 244–254
14 Zhang S, Xiong R, Sun F. Model predictive control for power management in a plug-in hybrid electric vehicle with a hybrid energy storage system. Appl Energy, 2017, 185: 1654–1662
15 Xu B, Rathod D, Zhang D, et al. Parametric study on reinforcement learning optimized energy management strategy for a hybrid electric vehicle. Appl Energy, 2020, 259: 114200
16 Feru E, Murgovski N, de Jager B, et al. Supervisory control of a heavy-duty diesel engine with an electrified waste heat recovery system. Control Eng Practice, 2016, 54: 190–201
17 Mansour C, Bou Nader W, Dumand C, et al. Waste heat recovery from engine coolant on mild hybrid vehicle using organic Rankine cycle. Proc Inst Mech Eng D-J Autom Eng, 2019, 233: 2502–2517
18 Kruijt K, Verbruggen F J R, Speetjens M F M, et al. Modeling and control of a waste heat recovery system for integrated powertrain design optimization. IFAC-PapersOnLine, 2019, 52: 598–603
19 Zhao C, Yin H, Yang Z, et al. Equivalent series resistance-based energy loss analysis of a battery semiactive hybrid energy storage system. IEEE Trans Energy Convers, 2015, 30: 1081–1091
20 Zhang Q, Ju F, Zhang S, et al. Power management for hybrid energy storage system of electric vehicles considering inaccurate terrain information. IEEE Trans Automat Sci Eng, 2017, 14: 608–618
21 Wieczorek M, Lewandowski M. A mathematical representation of an energy management strategy for hybrid energy storage system in electric vehicle and real time optimization using a genetic algorithm. Appl Energy, 2017, 192: 222–233
22 Buşoniu L, de Bruin T, Tolić D, et al. Reinforcement learning for control: Performance, stability, and deep approximators. Annu Rev Control, 2018, 46: 8–28
23 Zhang B, Hu W, Cao D, et al. Deep reinforcement learning-based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy. Energy Convers Manage, 2019, 202: 112199
24 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550: 354–359

25 Tan H, Zhang H, Peng J, et al. Energy management of hybrid electric bus based on deep reinforcement learning in continuous state and action space. Energy Convers Manage, 2019, 195: 548–560
26 Xiong R, Cao J, Yu Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. Appl Energy, 2018, 211: 538–548
27 Zou Y, Liu T, Liu D, et al. Reinforcement learning-based real-time energy management for a hybrid tracked vehicle. Appl Energy, 2016, 171: 372–382
28 Zou R, Fan L, Dong Y, et al. DQL energy management: An online-updated algorithm and its application in fix-line hybrid electric vehicle. Energy, 2021, 225: 120174
29 Lian R, Peng J, Wu Y, et al. Rule-interposing deep reinforcement learning based energy management strategy for power-split hybrid electric vehicle. Energy, 2020, 197: 117297
30 Wang X, Shu G, Tian H, et al. Dynamic analysis of the dual-loop Organic Rankine Cycle for waste heat recovery of a natural gas engine. Energy Convers Manage, 2017, 148: 724–736
31 Zhou M, Zhang Y, Wang X. Modeling and simulation of power assembly for signal-axle parallel hybrid electric vehicles. Electric Mach Control, 2019, 13: 36–40
32 Shu G, Li X, Tian H, et al. Alkanes as working fluids for high-temperature exhaust heat recovery of diesel engine using organic Rankine cycle. Appl Energy, 2014, 119: 204–217
33 Mikielewicz D, Mikielewicz J. A thermodynamic criterion for selection of working fluid for subcritical and supercritical domestic micro CHP. Appl Thermal Eng, 2010, 30: 2357–2362
34 Wang X, Shu G, Tian H, et al. Effect factors of part-load performance for various Organic Rankine cycles using in engine waste heat recovery. Energy Convers Manage, 2018, 174: 504–515
35 Gao Y, Wang X, Tian H, et al. Quantitative analysis of fuel-saving potential for waste heat recovery system integrated with hybrid electric vehicle. Int J Energy Res, 2020, 44: 11152–11170
36 Liu T, Wang E, Meng F, et al. Operation characteristics and transient simulation of an ICE-ORC combined system. Appl Sci, 2019, 9: 1639
37 Wang X, Shu G, Tian H, et al. Engine working condition effects on the dynamic response of organic Rankine cycle as exhaust waste heat recovery system. Appl Thermal Eng, 2017, 123: 670–681
38 Jensen J, Tummescheit H. Moving boundary models for dynamic simulation of two-phase flows. In: The 2nd International Modelica Conference. Oberpfaffenhofen, 2002. 18–19
39 Johan P, Paolino T, Olivier L. Improving the control performance of an organic Rankine cycle system for waste heat recovery from a heavy-duty diesel engine using a model-based approach. In: 52nd IEEE Conference on Decision and Control. Firenze, 2013. 10–13
40 Qi X, Luo Y, Wu G, et al. Deep reinforcement learning enabled self-learning control for energy efficient driving. Transp Res Part C-Emerg Tech, 2019, 99: 67–81
41 Pisu P, Rizzoni G. A comparative study of supervisory control strategies for hybrid electric vehicles. IEEE Trans Contr Syst Technol, 2007, 15: 506–518
42 Li Y, He H, Khajepour A, et al. Energy management for a power-split hybrid electric bus via deep reinforcement learning with terrain information. Appl Energy, 2019, 255: 113762
43 Wang X, Wang R, Jin M, et al. Control of superheat of organic Rankine cycle under transient heat source based on deep reinforcement learning. Appl Energy, 2020, 278: 115637