energies

MDPI

Article

Performance Assessment and Comparative Analysis of Photovoltaic-Battery System Scheduling in an Existing Zero-Energy House Based on Reinforcement Learning Control

Wenya Xu [1], Yanxue Li [1,2,*], Guanjie He [1], Yang Xu [1,3] and Weijun Gao [1,3]

[1] Innovation Institute for Sustainable Maritime Architecture Research and Technology, Qingdao University of Technology, Qingdao 266033, China; x2306590950@163.com (W.X.); gaoweijun@me.com (W.G.)
[2] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong 100872, China
[3] Faculty of Environmental Engineering, The University of Kitakyushu, Kitakyushu 808-0135, Japan
* Correspondence: liyanxue@qut.edu.cn; Tel.: +86-156-8997-8251

Abstract: The development of distributed renewable energy resources and smart energy management are efficient approaches to decarbonizing building energy systems. Reinforcement learning (RL) is a data-driven control algorithm that trains a large amount of data to learn control policy. However, this learning process generally presents low learning efficiency using real-world stochastic data. To address this challenge, this study proposes a model-based RL approach to optimize the operation of existing zero-energy houses considering PV generation consumption and energy costs. The model-based approach takes advantage of the inner understanding of the system dynamics; this knowledge improves the learning efficiency. A reward function is designed considering the physical constraints of battery storage, photovoltaic (PV) production feed-in profit, and energy cost. Measured data of a zero-energy house are used to train and test the proposed RL agent control, including *Q*-learning, deep *Q* network (DQN), and deep deterministic policy gradient (DDPG) agents. The results show that the proposed RL agents can achieve fast convergence during the training process. In comparison with the rule-based strategy, test cases verify the cost-effectiveness performances of proposed RL approaches in scheduling operations of the hybrid energy system under different scenarios. The comparative analysis of test periods shows that the DQN agent presents better energy cost-saving performances than *Q*-learning while the *Q*-learning agent presents more flexible action control of the battery with the fluctuation of real-time electricity prices. The DDPG algorithm can achieve the highest PV self-consumption ratio, 49.4%, and the self-sufficiency ratio reaches 36.7%. The DDPG algorithm outperforms rule-based operation by 7.2% for energy cost during test periods.

Keywords: reinforcement learning; reward design; battery storage; PV consumption; energy cost

Citation: Xu, W.; Li, Y.; He, G.; Xu, Y.; Gao, W. Performance Assessment and Comparative Analysis of Photovoltaic-Battery System Scheduling in an Existing Zero-Energy House Based on Reinforcement Learning Control. *Energies* **2023**, *16*, 4844. https://doi.org/10.3390/en16134844

Academic Editor: JongHoon Kim

Received: 1 May 2023
Revised: 16 June 2023
Accepted: 20 June 2023
Published: 21 June 2023

1. Introduction

Buildings account for 30–40% of total energy usage in many developed countries [1]. In the US, buildings consume 75% of the total national electricity, and building energy consumption contributed to 22% of national carbon emissions in China in 2021 [2]. In the context of buildings' high energy consumption, it is essential to promote the development of distributed renewable energy sources to decarbonize building energy systems [3,4]. Conventional buildings only rely on importing electricity from the public grid, while zero-energy houses (ZEHs) can use on-site energy sources as well as grid imports to meet the energy demand [5]. The basic ways to achieve ZEHs can be summarized as minimizing demand from buildings and maximizing consumption of variable renewable energy resources [6]. However, operational performances of existing distributed ZEHs usually perform poorly, including inadequate energy flexibility, low renewable energy

consumption ratio, and high operating costs. In addition, the issues of renewable energy sources consumption and real-time balancing between supply and demand have arisen. Renewable energy sources are intermittent and uncontrollable, and building user loads are time-varying. This mismatch between supply and demand limits the local consumption capacity and penetration of renewable energy [7,8]. These issues also pose challenges to the reliability of grid operation, especially while selling excess generated power to the grid [9]. It aggravates the fluctuation in the net load of the grid (i.e., the grid load minus renewable energy generation) [10]. The energy storage system can mitigate the mismatch between on-site energy supply and user demand [7]. In the renewable energy system of a ZEH, the proximity of renewable energy sources, load, and storage units makes it easy to coordinate the production, storage, and demand of energy at different times through real-time control, which promotes the local consumption of renewable energy. Results of previous works indicate that there is potential to increase cost-effectiveness [11] and local consumption performance [12] through real-time smart control. Real-time control for building energy systems has emerged as a promising alternative to improve operational cost-effectiveness and local renewable energy consumption performances [1,13].

1.1. Literature Review

1.1.1. Building Energy System Management

The rising penetration of distributed renewable energy resources has transformed a large number of buildings into active energy prosumers [6]. Activating the flexibility potential of demand-side resources is crucial to improve on-site PV consumption and cost-saving. Demand-side management (DSM) can be defined as the end-user behaviors that efficiently manage energy usage [14]. DSM can reduce energy consumption or cost through load shifting [6], load modulation [13,15], and other demand response (DR) strategies. Chia-Shing Tai demonstrated the potential for grid load shifting by implementing DSM strategies [16]. Additionally, researchers have focused on achieving flexible DR, particularly through controlling heating, ventilation, and air conditioning (HVAC) loads [17,18]. Max Bird et al. proposed HVAC control based on a model predictive control (MPC) algorithm. Results show the proposed approach can achieve an improvement of 1.7% cost saving while ensuring thermal comfort compared to the baseline [19]. Rui Tang et al. achieved a 19–22% energy reduction by modulating the demand power of a building's chillers to improve indoor thermal comfort [20].

In recent years, the academic community has placed considerable emphasis on optimizing control actions in building energy systems to participate in DSM. Prateek Munankarmi et al. controlled building loads, PV, and battery systems based on a home energy management system to minimize utility bills while maintaining occupant comfort. The bill of a ZEH pays approximately 22% less in the simulated results [21]. Lissy Langer et al. utilized real-time control strategies to optimize the operation of residential building energy systems with PV and heat storage. The energy systems achieved self-sufficiency at 79% and 80%, respectively [22,23]. Moreover, a long-term study conducted by Kun Zhang et al. demonstrated that effectively controlling the building energy system could lead to annual electricity cost savings of 12% and a 34% reduction in peak load demand while maintaining indoor thermal comfort [12].

1.1.2. Challenges of Building Energy System Control

Rule-based strategies have become popularly used for real-time control of building energy systems. Jafari et al. proposed a hybrid approach consisting of rule-based real-time controllers and dynamic planning-based optimization techniques for managing building energy systems. Their research found that energy costs for building users dropped by an impressive 85% compared to a grid-based energy supply [24]. Though traditional rule-based control may be effective for some operational scenarios, it relies heavily on prior experience for parameter setting and requires deterministic control strategies specific to each building.

Additionally, the rule-based approach has difficulty handling the uncertainties in building energy systems.

MPC has been involved in the field of building energy system control [17,20]. MPC is a constrained optimal control strategy that uses a prediction model of the building energy system to predict its future behavior. The central feature of MPC is that it can forecast the future evolution of the system over a rolling timescale [25]. MPC can ensure optimal control actions based on the optimization goals using these forecasts. The prediction model can be physics-based (white box), hybrid (gray box), or data-driven (black box) [1]. White boxes use physics information to describe the building energy systems. They emphasize the requirement for the accuracy of the physical model. The accurate physical model highlights the requirement for expert knowledge of users [26]. Additionally, the building energy system environment is often complex and variable. MPC models cannot be dynamically adjusted to adapt to the changes in the environment, leading to the increasing inaccuracy of the physical model [27]. In the existing research, many prediction models of MPC are white boxes. A number of large commercial office buildings successfully operate a cloud-based real-time white box MPC in the practical aspects [26]. Black boxes are based on data-driven models combined with machine learning [28] or statistics. However, few models discuss the applications of MPC based on black boxes in building energy systems [29]. Because of these issues, MPC algorithms for building-distributed energy systems are limited to broad implementation.

1.1.3. Reinforcement Learning

With the development of artificial intelligence techniques, the data-driven RL approach has emerged as an attractive alternative to MPC for optimizing energy system operations [23,30]. Pre-determined strategy and accurate physical models are not required in RL algorithms. RL learns the optimal control policies from interactions with the environment and selects the actions based on a given reward mechanism [27]. The control policies can be adaptively adjusted based on the feedback from the environment and show an advantage in adapting to the stochastic environment changes [31]. Table 1 summarizes different applications of RL algorithms in managing energy system operations. RL algorithms can be divided into model-based and model-free RL algorithms. Model-free RL agents learn the optimal policies through continuous experimentation without using prior knowledge regarding system dynamics. Enormous amounts of data and time are necessary for high-quality convergence effectiveness [27]. Model-free algorithms perform well in an environment where the dynamics change stably. However, real-world data with stochastic patterns generally lead to an unstable training process. Model-based algorithms take actions obtained based on a simple environment model, leading to low requirements for data and better generalization of convergence [31].

Table 1. Relevant review works of energy system operation based on RL.

Ref. &Year	Algorithm	Object	Model	Control Action	Reward	Purpose
Sara Abedi et al., 2022 [32]	*Q*-learning	Residential energy management	Model-free	Charge and discharge power of battery	Energy cost drawn from the grid	Storage efficiency, energy cost saving
Mohammed H. Alabdullah et al., 2022 [33]	DDQN	Microgrid energy management	Model-free	Power of the generation system, charge and discharge power of the storage system	Operational costs and penalty for violating power flow constraints	Calculation efficiency and operational cost saving
Rendong Shen et al., 2022 [34]	DDQN	Building energy management	Model-free	Airflow ratio, battery state of charge	Energy cost, unconsumed amount of renewable energy, thermal comfort	Equipment control considering comfort
Zhiqiang Wan et al., 2018 [35]	DQN	Residential energy management	Model-free	Electricity from grid	Energy cost and penalty for violating the allowed minimum SOC	Power cost saving
Sepehr Sanaye et al., 2021 [23]	DQN	Residential area energy management	Model-based	Load of combined heat and power generation, heat from gas water heater	Total operational cost at current time interval	Energy cost saving, self-sufficiency ratio
Alexander Dreher et al., 2022 [36]	PPO	Wind power and Hydrogen storage system	Model-based	Electrolyzer power	Electrolyzer applicable operational set point, power demand of the hydrogen compressor, amount of hydrogen, and natural gas utilized	Opportunity cost
Samir Touzani et al., 2021 [30]	DDPG	HVAC system and battery system	Model-free	Air supply temperature and flow, charging and discharge power of battery	Energy purchased from the grid/Energy cost, the penalty for the violation of the temperature comfort zone, and battery physical limits	Energy cost saving considering thermal comfort
Yuan Gao et al., 2022 [37]	DDPG& TD3	Office building energy management	Model-free	Continuous control of biomass generator, PV and battery	Amount of electricity purchased, and penalty for violating the given minimum and maximum SOC	Off-grid operation and battery safety

Most researchers have been attempting to optimize objectives (e.g., user cost reduction, renewable energy consumption, user comfort, and load flexibility) by implementing a single approach, such as the rule-based method, *Q*-learning algorithms, DQN, and other regulation strategies. Thus, it is necessary to analyze the strengths and weaknesses of each RL algorithm in detail. This emphasizes the importance of further discussion and investigation to compare the effectiveness of different RL algorithms in energy system operation.

1.2. Research Contributions

The main contributions of this work are summarized as follows:

- We formulate the operation problem of a hybrid energy system in ZEHs as Markov decision processes (MDP). Additionally, we adopt model-based RL algorithms to optimize the operation problem and design an appropriate reward function to guide the energy system operation. The proposed reward function considers the physical constraints of the battery and the energy cost.
- Measured data of a ZEH are applied to train and test the RL agents. The proposed approaches achieve outstanding operation performances utilizing a relatively small amount of data.
- The scheduling performances of *Q*-learning, DQN, and DDPG agents are compared. The study demonstrates the cost-effectiveness of the proposed RL control approach in different seasons.

2. Methodology

2.1. Reinforcement Learning

Reinforcement learning is highly regarded as an adaptable feedback controller that can respond to environmental changes. RL can achieve goal optimality in control problems while considering any uncertainty. Control problems are generally described as MDP [38]. The quaternion (s, a, r, π) is used to show the agent, and the agent can be defined through the following Equation (1):

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right], \quad (1)$$

where α represents the learning rate, which means the level of trustworthiness of the update section. γ represents the discount factor of the future Q value in the present. s represents the environment state, and a is the action that an agent can perform. r represents the defined reward equation according to demand. π represents the strategy of the agent, which determines the action to be taken at a given state and is the core of RL agents. The setting of parameters, such as α and γ, is an important part of a model. It influences the operation and convergence performance.

2.2. Q-Learning

The *Q*-learning algorithm defines the state-action value function based on target questions, state space, and control actions. This function is commonly referred to as the Q function. The agent employs the reward value to update the Q table by taking corresponding actions in different states. After several episodes, the agent can learn how to act to gain the highest accumulated reward and gain the optimum strategy collection to solve the target question. According to the Q table, the decision key involves selecting the action with the highest Q value. A significant feature of the *Q*-learning algorithm is its ease and computational efficiency. It usually uses discrete actions to solve problems concerning optimization and regulation with small action spaces and state spaces.

2.3. Deep Q Network

The *Q*-learning algorithm is suitable for solving low-dimensional state space questions. When the *Q*-learning algorithm is used to address practical problems, the issue of

"dimension disaster" occurs. The deep learning network has significant function-fitting capabilities. The Deep *Q*-Network (DQN) algorithm combines the strengths of *Q*-learning and deep learning algorithms by replacing the *Q* tables with deep learning networks. These networks take the state information of the temporal sequence as input and output the *Q* value of each action corresponding to the state. Actions are chosen based on the ε-greedy strategy [39,40].

The DQN algorithm introduces several innovations. (1) The use of evaluation and target networks. The evaluation network learns directly from the environment, and the target network maintains the stability of the target *Q* value over a certain period. It benefits mitigating the volatility of the model. (2) Experience replay, where a certain number of observations (s, a, r, s') are stored. Then, a batch size of samples is randomly selected from the replay for training. The evaluated *Q* value, target *Q* value, and reward are used to calculate losses for adjusting the evaluation network weights by feedback. The replay allows the network to obtain the weights to suit more scenarios. The loss, which is the squared error between the target *Q* value and the evaluated *Q* value, is calculated as follows in combination with Equation (1).

$$Q_{target} = r_{t+1} + \gamma \max_{a} Q(s_{t+1}, a), \tag{2}$$

$$Q_{eval} = Q(s_t, a_t|\theta), \tag{3}$$

$$\text{Loss} = E\Big[\left(Q_{target} - Q_{eval}\right)^2 \Big], \tag{4}$$

where Q_{target} represents the target *Q* value and Q_{eval} represents the evaluated *Q* value. θ represents the weights of the evaluated network. *Loss* represents the loss.

2.4. Deep Deterministic Policy Gradient

The deep deterministic policy gradient (DDPG) algorithm integrates the advantages of actor–critic algorithms in RL based on the DQN. This method allows the agent to obtain an optimal policy based on continuous action control [41]. The DDPG algorithm consists of four convolutional neural networks: actor network, critic network, target actor network, and target critic network. The actor network provides a deterministic action in the current state s based on the strategy, and the critic network displays the *Q* value of the action to be performed in the current state based on the value. The target actor network and the target critic network output the action a' and *Q* value Q' corresponding to the next state s', respectively. In contrast to the DQN evaluation network, which outputs the *Q* values of all discrete actions in the current state, the DDPG actor–critic network shows the *Q*-value of a deterministic action in the current state. Equations (5)–(8) provide the calculation of the loss parameter for updating the weight of the actor network and the loss parameter for updating the weight of the critic network.

$$a_i = \mu(s_i|\theta^{\mu}), \tag{5}$$

$$loss\_actor = \frac{1}{batch\ size} \sum_{i} Q(s_i, a_i), \tag{6}$$

$$y_i = r_i + \gamma Q'(s_{i+1}, a_{i+1}), \tag{7}$$

$$loss\_critic = \frac{1}{batch\ size} \sum_{i} (y_i - Q(s_i, a_i))^2, \tag{8}$$

where *batch size* denotes the number of samples from the replay, $1 \le i \le batch\ size$. a_i, s_i, and r_i are the action, state, and reward of a sample, respectively. Θ^{μ} represents weights

of actor network μ. Equation (5) represents the actor network μ and calculates the output action under the state s_i. Q is the Q value corresponding to a_i and s_i, and y_i is the Q value of the action taken in the current state. *Loss_actor* is the loss parameter of the actor network, and *loss_critic* is the loss parameter of the critic network.

3. Experiment Design

3.1. Reinforcement Learning Agent Design

This paper focuses on the energy system of ZEHs, comprised mainly of PV generation, a battery storage system, and user demand connected to the grid [34,35], as shown in Figure 1. Real-time electricity prices play an essential role in encouraging DR, which helps improve electricity consumption behavior in buildings. Consequently, the energy system is regulated based on real-time electricity prices to encourage effective interaction between buildings and the grid, ultimately aiming to save energy costs [42]. The control problem is described using the MDP. Figure 2 depicts the regulation process for a building's renewable energy system based on RL. The following sections introduce the details of our proposed data-driven, model-based RL algorithms.

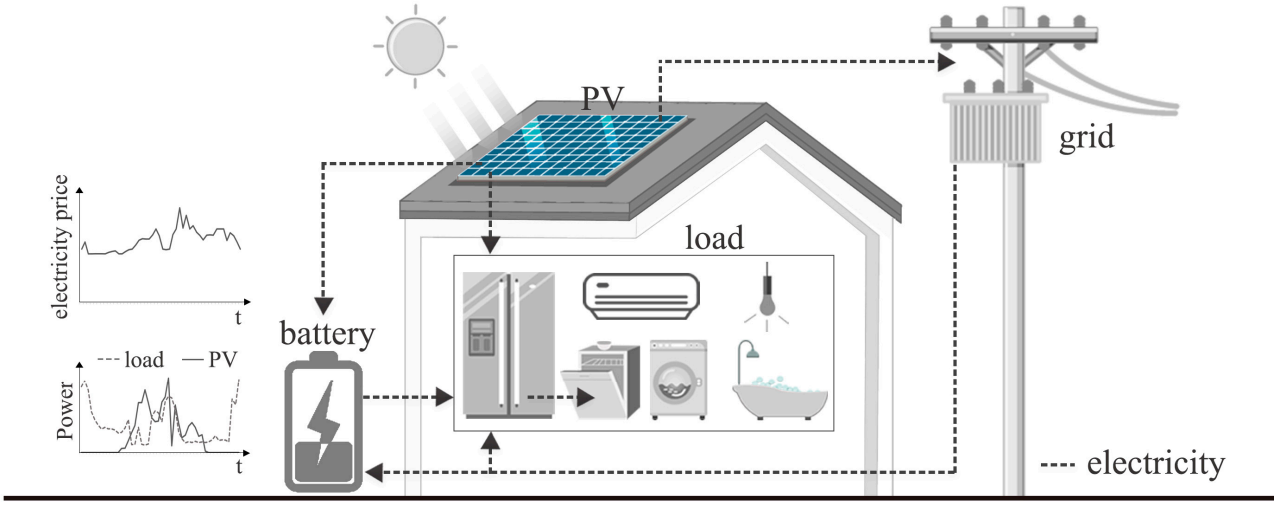

Figure 1. Structure of the hybrid energy system in ZEH.

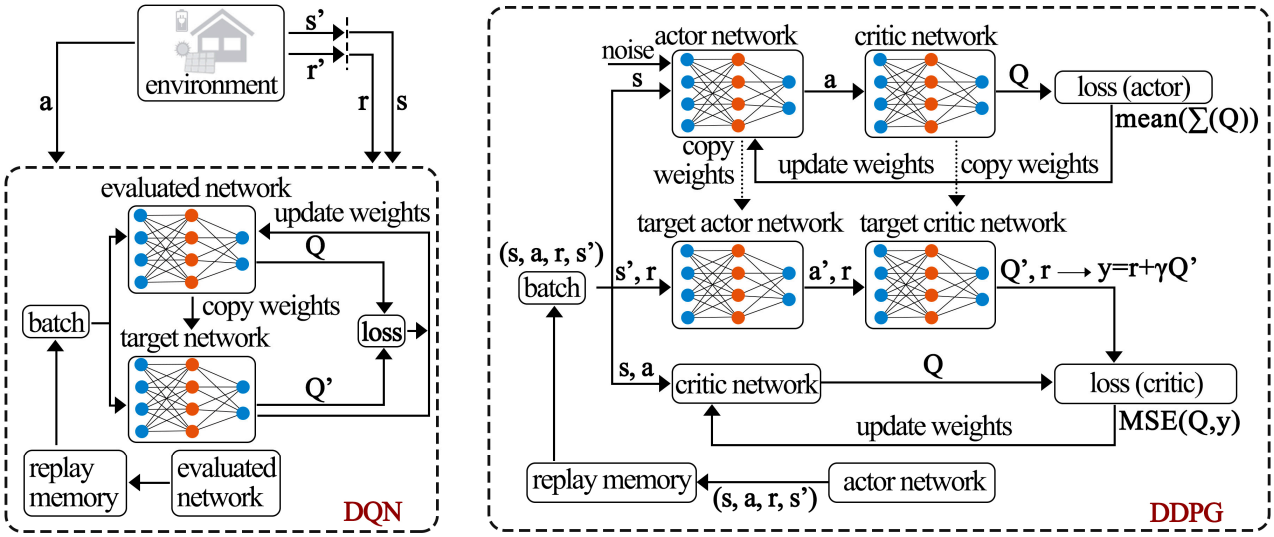

Figure 2. Schematic diagram of building renewable energy system regulation based on deep RL.

3.1.1. System State

The energy system model uses PV generation (kW), power demand (kW), state of charge of the battery, and real-time electricity prices as the RL observations. Thus, the state space is defined as follows:

$$s = \left[P_{pv}(t), P_{load}(t), SOC(t), R_{grid}(t)\right], \tag{9}$$

where t means at current time t, and $P_{load}(t)$ represents the user load at time t. Similarly, $P_{pv}(t)$ represents the PV generation power, and $R_{grid}(t)$ denotes the real-time electricity prices (JPY/kWh). $SOC(t)$ is the state of charge (SOC), and the maximum and minimum values of SOC are stable, SOC_{min} = 20% and SOC_{max} = 95%. As it is known, the Q-learning model cannot solve high-dimensional state problems. Therefore, the state components are divided into 10 discrete grids based on their maximum and minimum values at equal intervals. The DQN and DDPG model can handle the continuous state problems by neural networks instead of Q tables. Thus, the state components do not need to be discretized.

3.1.2. Control Action

In this work, the control action is the power charging and discharging activities. Action space is described as:

$$a = [P_{ba}(t)|g], \tag{10}$$

$$-P_{maxdch} \leq P_{ba}(t) \leq P_{maxch}, \tag{11}$$

where g represents the discretized granularity, g = 0.25, $P_{ba}(t)$ represents the discrete battery charge/discharge power at time t, P_{maxch} and P_{maxdch} are the maximum charge and discharge power (kW), $P_{maxch} = 0.3 \times E_{cap}$, and $P_{maxdch} = 0.25 \times E_{cap}$. The battery's charge/discharge power limit satisfies Equation (11). To avoid a local optimum solution, the algorithm uses the more common ε greedy strategy to ensure that each action is potentially chosen. The discretized granularity affects the speed of agent training and has little effect on the quality of learning. DQN and QL agents can both handle discrete action control problems. In this study, the discrete granularity settings are consistent. The control action of the DDPG agent is continuous, so discrete is unnecessary.

3.1.3. Reward Function

The RL agent can obtain the reward after performing the actions. The RL agent's learning goal is to maximize the accumulative reward. The reward can be expressed in certain values. The reward function guides the agent in optimization policy. Additionally, the reward function decides the value of a reward that the agent can obtain through taking each action. The value of the reward influences the learning direction of an agent.

The optimization of a ZEH energy system aims to improve on-site PV consumption and minimize energy costs. Furthermore, the safe operation of the battery needs to be considered. Thus, the reward function needs to involve the battery's physical constraints and energy cost, primarily composed of revenue from feeding PV and grid import costs. In this work, the reward function consists of two main components: the per-interval average cumulative cost resulting from the grid transactions and the reward on account of battery actions. First, the per-interval average cumulative cost replaces the energy cost in each interval [43]. It reduces the magnitude of reward values and expands the influence of the actions in the previous intervals [37]. Then, the function provides the battery operation with some soft constraints. When the battery has actions in the safe range of SOC, a reward of 1 is provided. Moreover, when the SOC violates the limits of the maximum and minimum values, a punishment of -20 is given. However, a punishment of -100 is also provided if the battery has no action because the agent may not run the batteries to avoid punishment. The values of punishment -100 and -20 are selected by several convergence experiments. The value of punishment can offset the cumulative positive reward so that

the agent chooses the right actions in specific states [44]. The proposed reward function guides the agent in using on-site energy sources and reducing the purchased electricity from the grid. Additionally, it directs the agent to choose the actions in the safe range of *SOC*. The reward function can be expressed by Equations (12)–(14).

$$m_{pri}(t) = \frac{1}{T}\sum_{t=1}^{T} P_{gr}(t) \times R_{grid}(t) \times \Delta t, \tag{12}$$

$$r_{ba}(t) = \begin{cases} 1, \ if\ action\ \neq 0 \text{ and } SOC_{min} \leq SOC(t) \leq SOC_{max} \\ -100, \ if\ action = 0 \text{ and } SOC_{min} \leq SOC(t) \leq SOC_{max} \\ -20, \ if\ SOC(t) > SOC_{max} \text{ or } SOC(t) < SOC_{min} \end{cases}, \tag{13}$$

$$r(t) = -m_{pri}(t) \times a + r_{ba}(t) \times b, \tag{14}$$

where $m_{pri}(t)$ represents the cost from the grid transactions, which comprises revenue from PV sales and expenditure from grid purchases (Yen). $P_{gr}(t)$ denotes the power of buy or sell electricity (kW). If $P_{gr}(t)$ is positive, the system purchases electricity from the grid, and if it is negative, the system sells electricity to the grid. $r_{ba}(t)$ represents the reward and penalty the battery can receive solely by choosing its action, and $r(t)$ represents the combined reward through performing the selected action. The reward weights, denoted by a and b, are set at 0.0001 and 1 in this study, respectively. The weights make the cumulative reward of all steps less than 0 and close to 0 indefinitely, which benefits the convergence of training models.

3.1.4. Parameter Settings

Table 2 shows the values of parameters used in *Q*-learning, DQN, and DDPG agents. It is an important part of building RL models to achieve convergence and obtain optimal optimization results by adjusting hyperparameters continually. The settings of these parameters are chosen through continuous experiments and experiences. It is common to directly observe the trend of the reward convergence curve and refer to the existing literature. Other parameters keep default settings in the simulation environment. It is known that the proposed RL agents use the same values of parameters except for several special parameters.

Table 2. Parameters in *Q*-learning, DQN, and DDPG.

Parameter	*Q*-Learning	DQN	DDPG
Discount factor γ	0.9	0.9	0.9
Learning rate α	0.001	0.001	0.001
ε-greedy initial value	0.9	0.9	0.9
ε-greedy decay factor	0.9	0.9	0.9
ε-greedy minimum value	0.01	0.01	0.01
Replay memory capacity	-	100,000	100,000
Batch size	-	32	32
Target net update steps	-	300	300
Number of hidden layers	-	2	2
Number of hidden units	-	(64, 64)	(64, 64)
Soft update coefficient	-	-	0.01

3.2. Rule-Based Operation

The rule-based method involves formulating regulated rules based on extensive experience, providing an operational strategy for optimizing energy systems [16]. In this work, the rule-based operation mainly aims to control the charge/discharge power of a battery. It is known that the load peaks of ZEHs shift to the periods of 3:00–6:00 and 19:00–22:00. An amount of PV generation is surplus over the daytime. Thus, the control

rules are as follows: First, the battery discharges accordingly with the load demand at 3:00–6:00 and 19:00–22:00. Additionally, PV energy is charged at 6:00–19:00.

3.3. Data Resource and Evaluation Method

In this work, we consider the energy system of a real ZEH located in Kyushu, Japan. We collected energy system data for the house over two distinct periods: from 1 January 2021 to 28 February 2022, and from 1 March 2018 to 28 August 2018. The collected data include load, PV generation, PV power sold, and power purchased from the grid, among other variables. Furthermore, the work uses the day-ahead market prices of Kyushu as real-time electricity prices with certain real-time and flexible features. The prices have been preprocessed, as shown in Figure 3. A 5-kWh battery with an 85% charge and discharge efficiency was used in the simulation. Additionally, the simulation incorporates a maximum of 1.5 kW charge power and a maximum of 1.25 kW discharge power and ensures that the SOC remains within the range of 20–95%. It is important to note that the simulation was conducted with a Python environment. Furthermore, the study was able to more accurately quantify the optimization and regulation of PV consumption by establishing quantitative relationships between distributed PV and user load. Several common assessment parameters were utilized; the details can be found in Table 3.

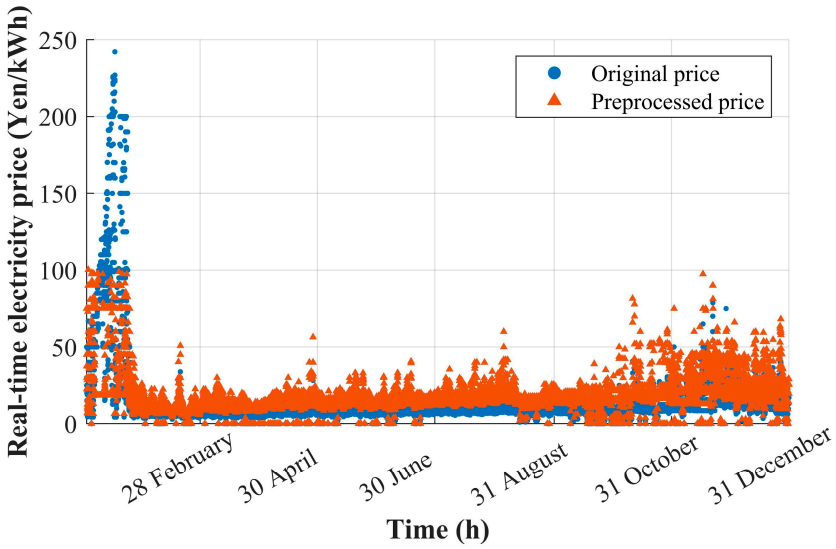

Figure 3. Real-time electricity prices.

Table 3. Common assessment parameters for evaluating the ZEH energy system.

Variables	Formula	Definition
Self-sufficiency ratio [14,45–47]	$\frac{\sum P_{pv-self\text{-}consumed}(t)}{\sum P_{load}(t)} \times 100\%$	The ratio of renewable power consumed by the user to total load demand.
Self-consumption ratio [14,45,46]	$\frac{\sum P_{pv-self\text{-}consumed}(t)}{\sum P_{pv}(t)} \times 100\%$	The ratio of renewable power consumed by the user to total renewable power.
Feed-in ratio [48,49]	$\frac{\sum P_{pv}(t) - \sum P_{pv-self\text{-}consumed}(t)}{\sum P_{pv}(t)} \times 100\%$	The ratio of renewable power sold to the grid to total renewable power.

4. Results and Discussions

4.1. Training Performance

The agents are trained using the dataset measured from 1 January 2021 to 31 December 2021. Then, the datasets measured from 1 January 2022 to 28 February 2022, and from

1 March 2018 to 31 July 2018 are used to test. The dataset is measured every half hour. The proposed *Q*-learning, DQN, and DDPG models are trained for 60 episodes to obtain the operation policy of the ZEH energy system. Each episode has 17,520 steps. It takes approximately 15 min per episode to finish the training of DQN and DDPG. The *Q*-learning agent takes 3 min per episode, offering a considerable amount of time savings. The optimization goal of the proposed agents is to minimize the operation cost of the ZEH energy system. Figure 4 illustrates the convergence of the cumulative operation cost in the training process. It can be seen that the proposed agents all keep learning and ultimately achieve a convergent policy. The training process goes through two stages, exploratory and convergence. Before 20 episodes, *Q*-learning and DDPG agents tend to explore actions, and the cumulative operation cost exhibits a reducing trend. After 20 episodes, *Q*-learning and DDPG agents achieve convergence. The DQN agent attains the convergence state and stabilizes at the highest value of operation cost after 10 episodes. Figure 4 shows that the cumulative operation costs decrease steadily before the convergence state in the training process. The *Q*-learning agent achieves the lowest cumulative operation cost at the end of the 60 episodes compared to the other proposed agents. The operation cost curve of the DQN agent displays evident fluctuations. The convergence reveals the proposed RL algorithm can obtain the available operation policy through continuous learning and exploratory. The training performance of the proposed RL agents also shows the effectiveness of the reward function.

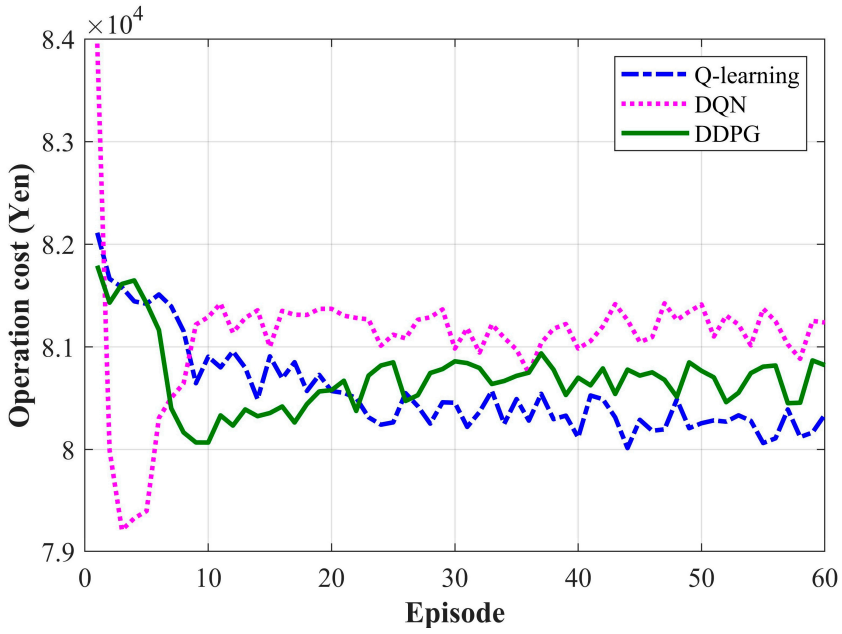

Figure 4. Cumulative operation costs performance in the training process.

4.2. Analysis of Testing Results

4.2.1. Effect of ZEH Energy System Scheduling

In this work, we analyzed the optimization of *Q*-learning, DQN, and DDPG agents compared to the rule-based operation. To explore and evaluate the optimization differences of the building energy system balance using the proposed RL agents, we analyzed two typical summer and winter weeks. The test results of energy system operation in a typical summer week are shown in Figure 5. Here, the battery importing is positive, and exporting is negative. For the grid exchange flows, the positive represents the system purchasing electricity from the grid, and the negative represents the system selling electricity to the grid. It is clear that electricity prices are at their highest during 10:00–14:00 when PV energy production is at its peak. The rule-based model uses this surplus PV electricity to charge for meeting user load demand at the specified times of the day. However, the models based on *Q*-learning, DQN, and DDPG demonstrated their flexibility by selling surplus

PV electricity and discharging to sell to the grid according to price fluctuations. During 23:00–5:00, the models based on DQN and DDPG usually satisfied the load and charge from the grid when the real-time electricity prices were low. After 20:00, the proposed optimization models regulated battery discharging to satisfy the building demand as PV generation levels were near zero. When demand peaked during 3:00–5:00, rule-based battery charging or discharging followed the established rules. In this typical summer week, the DQN and the *Q*-learning agents showed similar learning effects. However, the DDPG agent performed well in learning the ZEH energy system scheduling policy. The work aimed to improve PV generation local consumption and realize the minimization of energy costs. The DDPG agent used PV power to charge with relatively high power before the electricity prices peaked. Additionally, it discharged timely for the users' demand when electricity prices were relatively high.

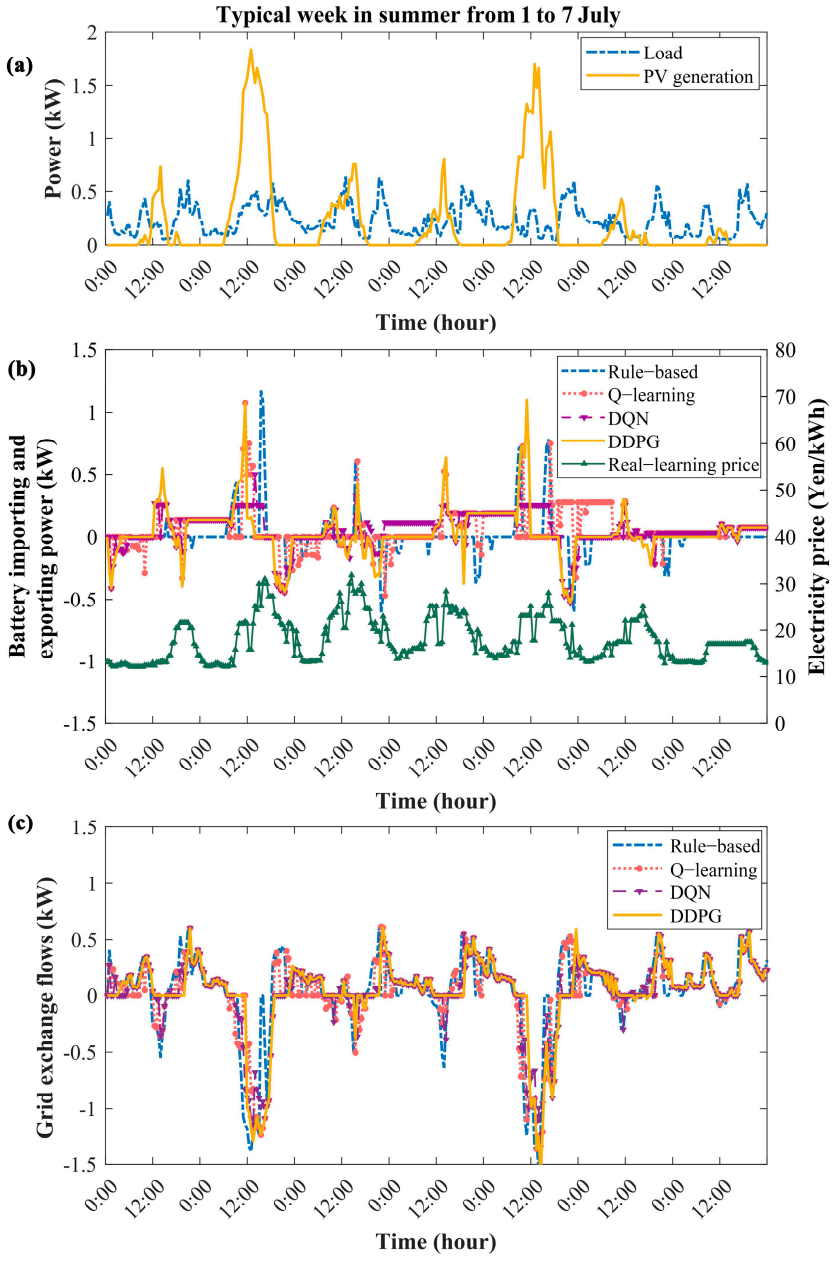

Figure 5. Operation of building energy system in a typical week of summer. (**a**) Operation of a ZEH l energy system; (**b**) Battery importing and exporting power; (**c**) Grid exchange flows.

The building energy system responses in a typical winter week are shown in Figure 6. It is known that PV generation decreases significantly in winter thanks to sunlight attenuation, and the load increases due to heating and hot water needs, etc. As a result, PV generation is insufficient to charge the batteries and meet the load, necessitating the purchase of additional electricity from the grid. The trends of grid transactions across all four models were consistent for the typical week. Real-time prices revealed that the tariff was lower during periods with high intra-day PV generation and increased when PV power was low or non-existent. In addition, the tariff for the first four days of this winter week was lower than that for the last three days. The results show that the building energy system charged continuously during 23:00–5:00 of the first four days despite PV generation's zero output. Although the reward function considered giving a certain penalty to the proposed agents when the batteries do not work, these situations still arise in the energy system operation during winter periods.

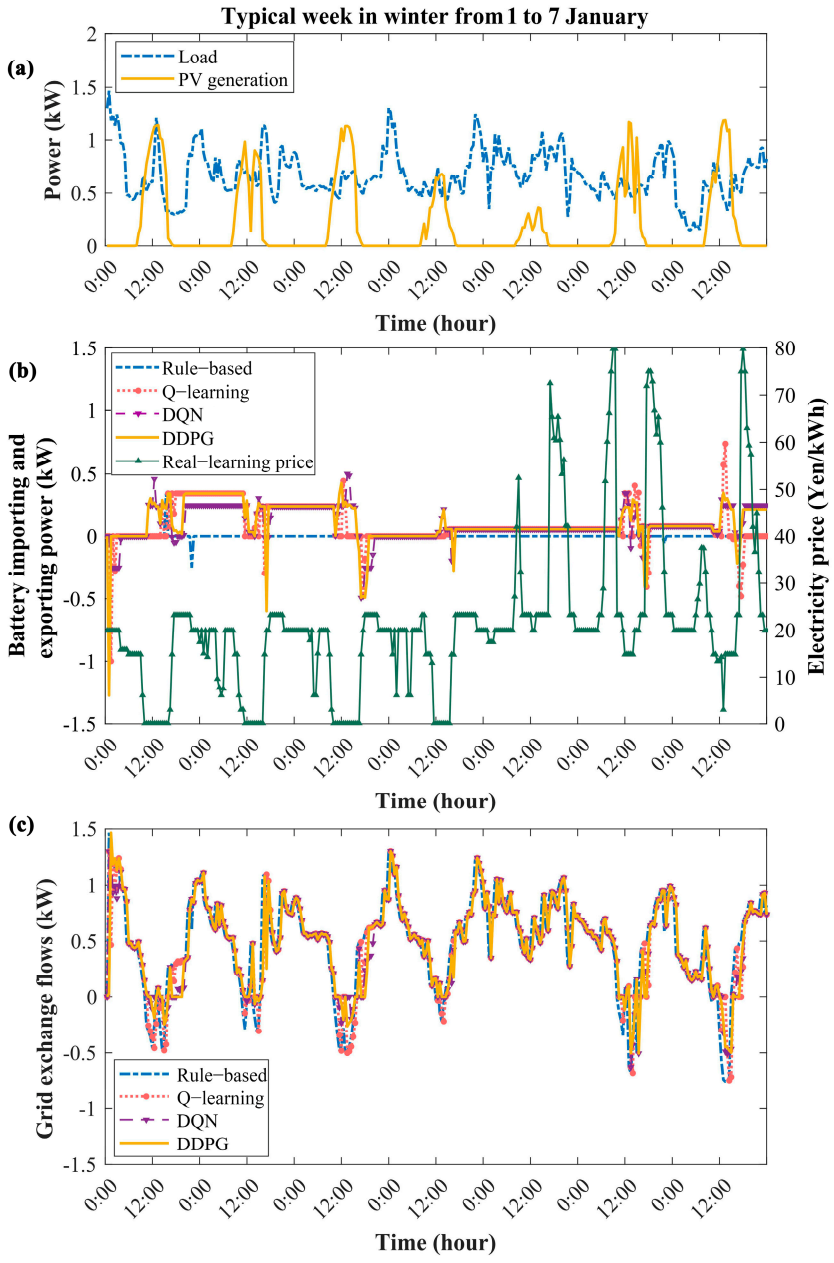

Figure 6. Operation of building energy system in a typical week of winter. (**a**) Operation of a ZEH l energy system; (**b**) Battery importing and exporting power; (**c**) Grid exchange flows.

It is noted that *Q*-learning and DQN typically select safer actions, rarely resulting in high-power charging or discharging in energy system operation. On the other hand, the DDPG model utilizes high-power actions more frequently. The battery in the model based on *Q*-learning tends to keep to low powers as actions, and its charging and discharge powers are greatly impacted by real-time electricity prices. This reveals the operation strategy based on the *Q*-learning algorithm obtains more flexibility. The proposed models obey the limits of the battery in their optimization. The agents learn the regularity of the ZEH energy system better. It is helpful for agents to judge if the operating environment stays within safe range though taking high-power actions and predicting the next steps. Thus, the DQN and *Q*-learning agents utilize relatively few high-power actions as they make decisions under their acquired strategies.

In addition, the paper highlights the effect of the proposed RL agents on load shifting. This work chose January, May, and July as winter, midseason, and typical summer months, respectively. Figure 7 shows the ZEH average net load curve using different optimization methods over three typical months. The net load is the electricity required to satisfy the demand not fulfilled by renewable energy, which is purchased from the grid by users. The net load peaks of this user can be observed during 2:00–6:00 and 19:00–23:00 over the three typical months. The highest net load peaks arose in operations based on *Q*-Learning among the proposed RL agents, and the agents all performed better than the rule-based model. In total, the building energy system model based on DDPG achieved the best peak-reducing results, effectively alleviating the electricity peak period load demands of users.

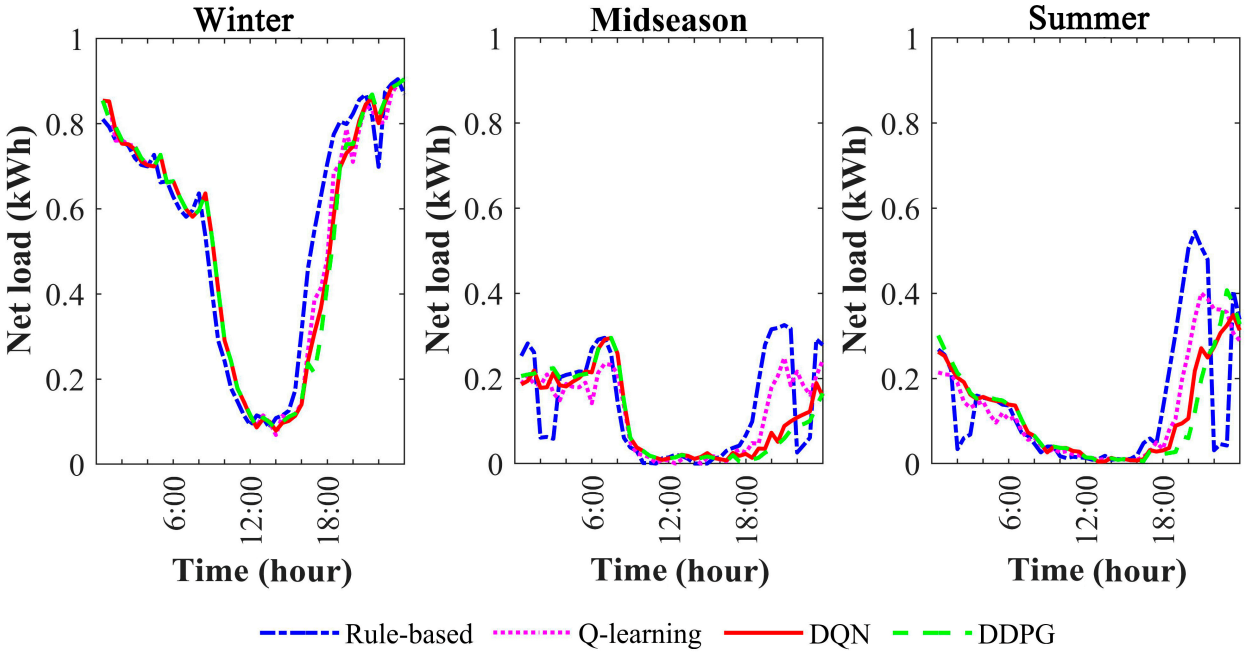

Figure 7. Average net load in the typical months.

4.2.2. PV Self-Consumption

In the simulation environment of Python, the work presents the optimization model based on *Q*-learning, DQN, and DDPG. The optimization results of PV consumption in building energy system operation are shown in Table 4. The system based on DDPG shows an outstanding self-consumption ratio of 49.4%, outperforming the compared rule-based result of 43.3%. The self-consumption levels with *Q*-learning and DQN methods are approximately the same at 47%. Compared to the self-consumption ratio of 35.8% without building energy system optimization, the proposed RL agents show a significant improvement in PV consumption ability. This parallel improvement in self-sufficiency level is mirrored in the increased PV self-consumption. More PV local consumption equates to

fewer feed-in PV generation instances. The DDPG method shows a decrease of 9.6%, while the DQN and *Q*-learning methods reduce by 8.0% and 5.0%, respectively, compared to the rule-based method.

Table 4. Operation results of optimization based on rule-based, *Q*-learning, DQN, and DDPG over the seven-month period tested.

Algorithm	Self-Consumption Ratio	SELF-Sufficiency Ratio	Feed-in Ratio
Rule-based	43.3%	32.2%	53.7%
Q-learning	47.4%	35.2%	48.7%
DQN	46.7%	34.7%	45.7%
DDPG	49.4%	36.7%	44.1%

By assessing the self-consumption ratio and feed-in ratio in Table 4, the PV generation loss can be calculated accordingly. It is revealed that the model based on DQN incurs the highest PV loss, 7.6%, while the battery charging and discharging in the *Q*-learning model results in a PV loss of 3.9%. On the other hand, the rule-based model produces the smallest loss, 3.0%, in the electricity schedule. The 5 kWh battery capacity, in combination with the proposed RL algorithms and the rule-based method, serves to reach the maximum energy self-sufficiency ratio of 36.7% and PV self-consumption ratio of 49.4% in the ZEH. This presents that the ZEH energy system still has much potential to consume PV locally. As the battery capacity is limited, the local consumption level is primarily determined by the mismatch between the load demand generated by the operation of heat pumps, air conditioners, etc., and PV generation.

Table 5 shows the PV consumption of the ZEH energy system in typical winter, summer, and spring-to-summer months. It can be seen that more than 70% of PV electricity in winter is used locally. Consequently, increasing PV electricity feed-in in the grid challenges its stability. The model based on DDPG is found to have the highest self-consumption ratio in typical months, even reaching 80% in the typical winter month, while the model based on DQN outperforms *Q*-learning.

Table 5. Operation results of optimization based on rule-based, *Q*-learning, DQN, and DDPG in typical months.

Variables \ Typical Month		Winter	Midseason	Summer
Self-consumption ratio	Rule-based	73.0%	27.2%	44.2%
	Q-learning	79.2%	31.5%	46.4%
	DQN	78.1%	31.7%	47.4%
	DDPG	80.0%	34.0%	48.6%
Self-sufficiency ratio	Rule-based	19.9%	44.3%	56.8%
	Q-learning	21.6%	51.3%	59.6%
	DQN	21.3%	51.7%	60.9%
	DDPG	21.8%	55.5%	62.4%
Feed-in ratio	Rule-based	26.2%	69.4%	51.7%
	Q-learning	15.9%	64.2%	48.9%
	DQN	12.5%	61.2%	46.4%
	DDPG	10.1%	60.3%	45.8%

4.2.3. Energy Cost Saving

Table 6 summarizes the cumulative operation cost resulting from simulating the operation of the ZEH energy system with test data. Negative values represent the revenue obtained by the users, while positive values represent expenditures due to the electricity schedule. Firstly, when using the rule-based results as a benchmark for comparison, the model based on DDPG enables an economization of 7.2% in the operation cost of the ZEH energy system over the seven-month period tested. Additionally, the operation cost based

on the DQN algorithm is found to economize 1.1% more than the *Q*-learning algorithm. Furthermore, the proposed models are all able to obtain relatively high revenue in the operation of a summer month among the typical months. It is worth noting that the revenue based on DDPG is 45.5% higher in the typical summer month. Moreover, the transaction between the grid and battery is significantly increased due to the decreasing PV electricity and the rising demand. These charging and discharging actions result in a loss, further increasing the operation cost, which explains why the operation cost based on DDPG only saves 7.2%.

Table 6. Cumulative operation cost in typical months.

Typical Month / Operation Cost (Yen)	Winter	Midseason	Summer	Total Cost
Rule-based	23,076	−3053	−2888	43,301
Q-learning	22,749	−2995	−3712	41,374
DQN	22,616	−2987	−3715	40,868
DDPG	22,528	−3040	−4201	40,201

5. Conclusions

This work applied RL algorithms, including *Q*-learning, DQN, and DDPG, to optimize the operations of the hybrid energy system in ZEHs. The optimization goal was formulated to minimize the energy operation cost in the real-time electricity market while improving renewable electricity self-consumption. The reward function was properly designed to guide the learning process of RL agents. The *Q*-learning, DQN, and DDPG agents were trained and tested based on the measured data. Results compare the dispatch performances of the proposed RL agents in various situations in terms of dynamic PV self-consumption and energy cost of test periods. The main conclusions can be drawn as follows:

The proposed RL algorithms showed good performances in achieving the supply and demand balance of building energy systems with uncertainty in various situations. The DDPG agent showed the best learning results, which can control the battery by taking high-power actions in typical weeks. The performances of the agents proved the effectiveness of the designed reward function. In detail, the optimization based on the DDPG algorithm outperformed in shifting load in terms of cost saving. Additionally, the *Q*-learning agent showed more flexible action control of the battery with the fluctuation of real-time electricity prices than the DQN algorithm.

The proposed RL algorithms increased the local PV electricity consumption by managing the behavior of a 5 kWh battery. Compared with the rule-based schedule results, the DDPG agent achieved the highest self-consumption ratio of 49.4% and self-sufficiency ratio of 36.7% in the testing results. The DQN agent performed better than the *Q*-learning agent in improving PV electricity consumption. However, the *Q*-learning agent achieved a lower energy loss of 3.9% caused by the charging/discharging of the battery during the test periods.

The DDPG agent achieved the lowest energy cost over the seven-month period tested. The energy cost saving was 7.2% higher than rule-based control, realizing better economic performance. Then, the energy cost based on DQN showed better economic benefits than *Q*-learning control. Compared with the rule-based testing results, the revenue of ZEH energy system operation optimization based on DDPG was more than 45.5% in a typical summer month.

Future research may expand on our findings to improve the evaluation system considering energy flexibility and energy efficiency and use multi-agent algorithms for the optimization operation of ZEHs. Moreover, other storage, such as thermal storage and hydrogen storage, can be considered to cooperate with the grid-building interaction to increase the potential of renewable energy self-consumption in buildings.

Author Contributions: Conceptualization, W.X., Y.L. and W.G.; methodology, W.X., Y.L., Y.X. and W.G.; software, W.X., Y.L., G.H. and Y.X.; validation, W.X. and Y.L.; formal analysis, W.X., Y.L. and W.G.; investigation, W.X.; resources, G.H.; data curation, W.X.; writing—original draft, W.X.; Writing—review & editing, W.X., Y.L. and Y.X.; visualization, W.X. and G.H.; supervision, W.X., Y.L. and W.G.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the China National Key R&D Program 'Research on the Energy Efficiency and Health Performance Improvement of Building Operations based on Lifecycle Carbon Emissions Reduction', grant number 2018YFE0106100, the Shandong Natural Science Foundation 'Research on Flexible District Integrated Energy System under High Penetration Level of Renewable Energy', grant number ZR2021QE084, and the Xiangjiang Plan 'Development of Smart Building Management Technologies Towards Carbon Neutrality', grant number XJ20220028.

Data Availability Statement: Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

DDQN	Double deep Q network
DQN	Deep Q network
DR	Demand response
DSM	Demand-side management
HVAC	Heating, ventilation and air conditioning
MDP	Markov decision processes
MPC	Model predictive control
PV	Photovoltaic
RL	Reinforcement learning
SOC	State of charge
TD3	Twin-delayed deep deterministic policy gradient
ZEH	Zero-energy house

References

1. Li, H.; Wang, Z.; Hong, T.; Piette, M.A. Energy flexibility of residential buildings: A systematic review of characterization and quantification methods and applications. *Adv. Appl. Energy* **2021**, *3*, 100054. [CrossRef]
2. Hu, S.; Yan, D. *China Building Energy Use and Carbon Emission Yearbook 2021: A Roadmap to Carbon Neutrality by 2060*; Springer Nature: Singapore, 2022.
3. Niu, J.; Zhou, R.; Tian, Z.; Zhu, J.; Lu, Y.; Ma, J. Energy-saving potential analysis for a 24-h operating chiller plant using the model-based global optimization method. *J. Build. Eng.* **2023**, *69*, 106213. [CrossRef]
4. Li, Y.; Wang, Z.; Xu, W.; Gao, W.; Xu, Y.; Xiao, F. Modeling and energy dynamic control for a ZEH via hybrid model-based deep reinforcement learning. *Energy* **2023**, *277*, 127627. [CrossRef]
5. Zhang, W.; Yan, C.; Xu, Y.; Fang, J.; Pan, Y. A critical review of the performance evaluation and optimization of grid interactions between zero-energy buildings and power grids. *Sustain. Cities Soc.* **2022**, *86*, 104123. [CrossRef]
6. Wu, Y.; Wu, Y.; Guerrero, J.M.; Vasquez, J.C. Decentralized transactive energy community in edge grid with positive buildings and interactive electric vehicles. *Int. J. Electr. Power Energy Syst.* **2022**, *135*, 107510. [CrossRef]
7. Wu, Y.; Liu, Z.; Li, B.; Liu, J.; Zhang, L. Energy management strategy and optimal battery capacity for flexible PV-battery system under time-of-use tariff. *Renew. Energy* **2022**, *200*, 558–570. [CrossRef]
8. He, F.; Bo, R.; Hu, C.; Meng, X.; Gao, W. Employing spiral fins to improve the thermal performance of phase-change materials in shell-tube latent heat storage units. *Renew. Energy* **2023**, *203*, 518–528. [CrossRef]
9. Jin, X.; Xiao, F.; Zhang, C.; Chen, Z. Semi-supervised learning based framework for urban level building electricity consumption prediction. *Appl. Energy* **2022**, *328*, 120210. [CrossRef]
10. Tang, H.; Wang, S.; Li, H. Flexibility categorization, sources, capabilities and technologies for energy-flexible and grid-responsive buildings: State-of-the-art and future perspective. *Energy* **2020**, *219*, 119598. [CrossRef]
11. Eslami, M.; Nahani, P. How policies affect the cost-effectiveness of residential renewable energy in Iran: A techno-economic analysis for optimization. *Utilities Policy* **2021**, *72*, 101254. [CrossRef]
12. Zhang, K.; Prakash, A.; Paul, L.; Blum, D.; Alstone, P.; Zoellick, J.; Brown, R.; Pritoni, M. Model predictive control for demand flexibility: Real-world operation of a commercial building with photovoltaic and battery systems. *Adv. Appl. Energy* **2022**, *7*, 100099. [CrossRef]

13. Bay, C.J.; Chintala, R.; Chinde, V.; King, J. Distributed model predictive control for coordinated, grid-interactive buildings. *Appl. Energy* **2022**, *312*, 118612. [CrossRef]
14. Pinto, G.; Kathirgamanathan, A.; Mangina, E.; Finn, D.P.; Capozzoli, A. Enhancing energy management in grid-interactive buildings: A comparison among cooperative and coordinated architectures. *Appl. Energy* **2022**, *310*, 118497. [CrossRef]
15. Wang, Y.; Xu, Y.; Tang, Y. Distributed aggregation control of grid-interactive smart buildings for power system frequency support. *Appl. Energy* **2019**, *251*, 113371. [CrossRef]
16. Tai, C.-S.; Hong, J.-H.; Hong, D.-Y.; Fu, L.-C. A real-time demand-side management system considering user preference with adaptive deep Q learning in home area network. *Sustain. Energy Grids Netw.* **2021**, *29*, 100572. [CrossRef]
17. Serale, G.; Fiorentini, M.; Capozzoli, A.; Bernardini, D.; Bemporad, A. Model Predictive Control (MPC) for Enhancing Building and HVAC System Energy Efficiency: Problem Formulation, Applications and Opportunities. *Energies* **2018**, *11*, 631. [CrossRef]
18. Li, S.; Zhang, X.; Li, Y.; Gao, W.; Xiao, F.; Xu, Y. A comprehensive review of impact assessment of indoor thermal environment on work and cognitive performance-Combined physiological measurements and machine learning. *J. Build. Eng.* **2023**, *71*, 106417. [CrossRef]
19. Bird, M.; Daveau, C.; O'Dwyer, E.; Acha, S.; Shah, N. Real-world implementation and cost of a cloud-based MPC retrofit for HVAC control systems in commercial buildings. *Energy Build.* **2022**, *270*, 112269. [CrossRef]
20. Tang, R.; Fan, C.; Zeng, F.; Feng, W. Data-driven model predictive control for power demand management and fast demand response of commercial buildings using support vector regression. *Build. Simul.* **2021**, *15*, 317–331. [CrossRef]
21. Munankarmi, P.; Maguire, J.; Balamurugan, S.P.; Blonsky, M.; Roberts, D.; Jin, X. Community-scale interaction of energy efficiency and demand flexibility in residential buildings. *Appl. Energy* **2021**, *298*, 117149. [CrossRef]
22. Langer, L.; Volling, T. A reinforcement learning approach to home energy management for modulating heat pumps and photovoltaic systems. *Appl. Energy* **2022**, *327*, 120020. [CrossRef]
23. Sanaye, S.; Sarrafi, A. A novel energy management method based on Deep Q Network algorithm for low operating cost of an integrated hybrid system. *Energy Rep.* **2021**, *7*, 2647–2663. [CrossRef]
24. Jafari, M.; Malekjamshidi, Z. Optimal energy management of a residential-based hybrid renewable energy system using rule-based real-time control and 2D dynamic programming optimization method. *Renew. Energy* **2019**, *146*, 254–266. [CrossRef]
25. Morato, M.M.; Mendes, P.R.C.; Normey-Rico, J.E.; Bordons, C. LPV-MPC fault-tolerant energy management strategy for renewable microgrids. *Int. J. Electr. Power Energy Syst.* **2019**, *117*, 105644. [CrossRef]
26. Drgoňa, J.; Arroyo, J.; Cupeiro Figueroa, I.; Blum, D.; Arendt, K.; Kim, D.; Ollé, E.P.; Oravec, J.; Wetter, M.; Vrabie, D.L.; et al. All you need to know about model predictive control for buildings. *Annu. Rev. Control* **2020**, *50*, 190–232. [CrossRef]
27. Lee, H.; Kim, K.; Kim, N.; Cha, S.W. Energy efficient speed planning of electric vehicles for car-following scenario using model-based reinforcement learning. *Appl. Energy* **2022**, *313*, 118460. [CrossRef]
28. Chen, Z.; Xiao, F.; Guo, F.; Yan, J. Interpretable machine learning for building energy management: A state-of-the-art review. *Adv. Appl. Energy* **2023**, *9*, 100123. [CrossRef]
29. Gao, Y.; Matsunami, Y.; Miyata, S.; Akashi, Y. Model predictive control of a building renewable energy system based on a long short-term hybrid model. *Sustain. Cities Soc.* **2023**, *89*, 104317. [CrossRef]
30. Touzani, S.; Prakash, A.K.; Wang, Z.; Agarwal, S.; Pritoni, M.; Kiran, M.; Brown, R.; Granderson, J. Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency. *Appl. Energy* **2021**, *304*, 117733. [CrossRef]
31. Totaro, S.; Boukas, I.; Jonsson, A.; Cornélusse, B. Lifelong control of off-grid microgrid with model-based reinforcement learning. *Energy* **2021**, *232*, 121035. [CrossRef]
32. Abedi, S.; Yoon, S.W.; Kwon, S. Battery energy storage control using a reinforcement learning approach with cyclic time-dependent Markov process. *Int. J. Electr. Power Energy Syst.* **2021**, *134*, 107368. [CrossRef]
33. Alabdullah, M.H.; Abido, M.A. Microgrid energy management using deep Q-network reinforcement learning. *Alex. Eng. J.* **2022**, *61*, 9069–9078. [CrossRef]
34. Shen, R.; Zhong, S.; Wen, X.; An, Q.; Zheng, R.; Li, Y.; Zhao, J. Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. *Appl. Energy* **2022**, *312*, 118724. [CrossRef]
35. Wan, Z.; Li, H.; He, H. Residential Energy Management with Deep Reinforcement Learning. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7. [CrossRef]
36. Dreher, A.; Bexten, T.; Sieker, T.; Lehna, M.; Schütt, J.; Scholz, C.; Wirsum, M. AI agents envisioning the future: Forecast-based operation of renewable energy storage systems using hydrogen with Deep Reinforcement Learning. *Energy Convers. Manag.* **2022**, *258*, 115401. [CrossRef]
37. Gao, Y.; Matsunami, Y.; Miyata, S.; Akashi, Y. Operational optimization for off-grid renewable building energy system using deep reinforcement learning. *Appl. Energy* **2022**, *325*, 119783. [CrossRef]
38. Richard, S.; Sutton, A.G.B. *Reinforcement Learning: An Introduction*, 2nd ed.; Publishing House of Electronics Industry: Beijing, China, 2019.
39. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
40. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing Atari with Deep Reinforcement Learning. *arXiv* **2013**, arXiv:1312.5602.

41. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
42. Zhang, H.; Xiao, F.; Zhang, C.; Li, R. A multi-agent system based coordinated multi-objective optimal load scheduling strategy using marginal emission factors for building cluster demand response. *Energy Build.* **2023**, *281*, 112765. [CrossRef]
43. Feng, J.; Wang, H.; Yang, Z.; Chen, Z.; Li, Y.; Yang, J.; Wang, K. Economic dispatch of industrial park considering uncertainty of renewable energy based on a deep reinforcement learning approach. *Sustain. Energy Grids Netw.* **2023**, *34*, 101050. [CrossRef]
44. Chen, H.; Zhuang, J.; Zhou, G.; Wang, Y.; Sun, Z.; Levron, Y. Emergency load shedding strategy for high renewable energy penetrated power systems based on deep reinforcement learning. *Energy Rep.* **2023**, *9*, 434–443. [CrossRef]
45. Liu, J.; Yin, R.; Yu, L.; Piette, M.A.; Pritoni, M.; Casillas, A.; Xie, J.; Hong, T.; Neukomm, M.; Schwartz, P. Defining and applying an electricity demand flexibility benchmarking metrics framework for grid-interactive efficient commercial buildings. *Adv. Appl. Energy* **2022**, *8*, 100107. [CrossRef]
46. Bee, E.; Prada, A.; Baggio, P.; Psimopoulos, E. Air-source heat pump and photovoltaic systems for residential heating and cooling: Potential of self-consumption in different European climates. *Build. Simul.* **2019**, *12*, 453–463. [CrossRef]
47. Puranen, P.; Kosonen, A.; Ahola, J. Techno-economic viability of energy storage concepts combined with a residential solar photovoltaic system: A case study from Finland. *Appl. Energy* **2021**, *298*, 117199. [CrossRef]
48. Li, Y.; Gao, W.; Ruan, Y. Performance investigation of grid-connected residential PV-battery system focusing on enhancing self-consumption and peak shaving in Kyushu, Japan. *Renew. Energy* **2018**, *127*, 514–523. [CrossRef]
49. Pacudan, R. Feed-in tariff vs incentivized self-consumption: Options for residential solar PV policy in Brunei Darussalam. *Renew. Energy* **2018**, *122*, 362–374. [CrossRef]