Contents lists available at ScienceDirect

Applied Soft Computing

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

Dynamic pricing and energy management for shore side electricity in a port microgrid: A deep reinforcement learning approach

Chungkwon Oh[a], Ilkyeong Moon[b,c,*]

[a] Digital Twin Team, MICUBE Solution, 23 Hyoryeong-ro 55-gil, Seocho-gu, Seoul 06654, Republic of Korea
[b] Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[c] Institute of Engineering Research, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

HIGHLIGHTS

- We present a profitable model for operating shore side electricity in a port microgrid.
- We address the dynamic pricing and energy management problem.
- Reinforcement learning algorithms are proposed for maximizing the port's profit.
- A reward shaping technique based on the myopic optimization model improves the reinforcement learning algorithm's performance.
- Positive side effects are observed in terms of the utilization of SSE and the overall system profit.

ARTICLE INFO

Keywords:
Shore side electricity
Port microgrid
Dynamic pricing
Energy management
Reinforcement learning

ABSTRACT

This paper addresses the economic challenges associated with operating shore side electricity (SSE) in ports, a critical measure for reducing greenhouse gas emissions in the maritime industry. Although SSE offers significant environmental benefits, its widespread adoption is hindered by operational costs and the volatility of electricity prices. To ensure the sustainable operation of SSE, we propose a dynamic pricing and energy management strategy integrated within a port microgrid to maximize operational profit. We leverage an actor-critic reinforcement learning approach and propose a reward shaping technique based on a myopic algorithm to enhance performance and stability. Through several experiments, we show that the proposed algorithm improves the port's profit by approximately 4.4 % compared to the rule-based heuristic algorithm, while also exhibiting greater stability and robustness in the learning process. We also demonstrate that integrating SSE with port microgrids can increase SSE utilization and benefit the entire system.

1. Introduction

The maritime industry is responsible for more than 80 % of global trade volume, making it a critical part of the global supply chain [1]. However, it is also a significant source of global air pollution, emitting 1.076 billion tons of greenhouse gas (GHG) emissions annually, which accounts for 2.89 % of the world's total GHG emissions [2]. With growing concerns over environmental pollution, the maritime industry is also moving toward decarbonization. As a key participant in the industry, ports are implementing various environmental policies, one of which is shore side electricity (SSE, also known as cold ironing, onshore power supply, shore power, alternative maritime power, and other terms). SSE refers to supplying electricity from the shore to a vessel while it is berthed instead of having the vessel generate electricity using its auxiliary engines. It has significant environmental benefits compared to using vessel fuel to run the engines, reducing carbon dioxide emissions by 49 % and sulfur dioxide emissions by 69 % [3].

Due to these factors, many organizations and governments worldwide have implemented regulations or incentive policies to encourage SSE utilization. For example, the California Air Resources Board in the United States mandated that beginning in 2023, vessels berthed at ports must source at least 80 % of their required electricity from SSE. In response, ports within California, including the ports of Los Angeles, Long Beach, Oakland, and others, have supplied SSE, allowing berthed vessels

[*] Corresponding author at: Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea.
Email addresses: ock0206@snu.ac.kr (C. Oh), ikmoon@snu.ac.kr (I. Moon).

https://doi.org/10.1016/j.asoc.2025.114089
Received 10 April 2025; Received in revised form 30 September 2025; Accepted 7 October 2025

Available online 18 October 2025

to comply with these regulatory standards by using SSE. Meanwhile, the Shanghai Municipal Government in China offers incentives to vessels utilizing SSE, including tariff subsidies or priority use of berths. Currently, most of the SSE operations rely on such regulations and policies. Thus, numerous studies have been conducted on effective policies aimed at increasing the utilization rate of SSE [4–6]. However, despite such interventions, the global utilization rate remains low because of its low cost-effectiveness [7]. Ideally, vessels can gain economic advantages by using SSE at a lower price compared to using traditional vessel fuel, while ports benefit economically by selling SSE at a higher price than the wholesale electricity price. Nevertheless, the volatility of electricity prices and their high average price have been obstacles for both vessels and ports in bearing the initial and operational costs. Recently, rising vessel fuel prices, driven by the International Maritime Organization's fuel regulations, have incentivized vessels to use SSE [8]. However, from the port's perspective, it is still difficult to realize the economic benefits of operating SSE, so it remains a challenge that should be addressed to increase the utilization of SSE.

To address the above problem, Ahamad et al. [9] were the first to propose the concept of a port microgrid as a solution to reduce emissions, incorporating SSE as a key element. A microgrid is a small-scale power grid that includes technologies such as renewable energy sources and an energy storage system (ESS), aiming for efficient energy operation through intelligent control of generation and load [10]. It also ensures high reliability and efficiency because it operates in a grid-connected mode (connected to the main grid) during normal conditions. In practice, there has been significant movement toward establishing port microgrids, as ports are suitable for centralized energy management and installing facilities such as offshore wind power, solar power, and ESS. In line with this trend, several studies have focused on optimizing the operation of individual components within the microgrid to ensure their effective integration [11,12]. By using the ESS, ports can effectively respond to the volatility in real-time electricity prices and renewable energy generation, enabling the development of a profitable model for operating SSE. The key factor for operational efficiency is how effectively the port controls the balance between the demand and supply of electricity. The port can control electricity demand by adjusting the prices of SSE and manage supply by purchasing electricity from the main grid. Thus, considering the uncertainties in electricity prices and renewable energy generation, energy management and pricing for SSE are crucial for profitable SSE operation.

To the best of our knowledge, no research has addressed the above issue at the operational level. This paper aims to deal with the pricing and energy management problem for operating SSE in a port microgrid. Specifically, this study focuses on the real-time electricity market, where energy management is more challenging than in the day-ahead market due to the uncertainty of electricity prices. In the day-ahead market, where hourly electricity prices are established daily, the uncertainty is lower, and relatively accurate modeling is possible, making model-based approaches more appropriate. However, applying a model-based approach in the real-time market requires the unrealistic assumption that future electricity prices can be predicted accurately. Additionally, as the time horizon extends and the problem size increases, the computational time required for model-based approaches increases significantly. Instead, we adopt reinforcement learning (RL) approaches that can be applied in model-free situations without relying on the unrealistic assumption. With RL methods, it is possible to make dynamic decisions based on real-time data within a short computational time.

We propose a Markov decision process (MDP) formulation and actor-critic RL algorithms to solve the pricing and energy management problem of maximizing the port's profit. We also present numerical experiments to analyze the algorithms' results and performance. The main contributions of this study are presented as follows:

1. We present a profitable model for operating SSE in ports and propose an RL algorithm with a reward shaping technique for pricing and energy management strategies. The proposed algorithm will enable ports to bear initial and operational costs by enhancing profitability. Ultimately, it serves as a foundational framework for enhancing the sustainability of operating SSE.
2. We validate the performance of the algorithm using real-world data, demonstrating its application in practical situations. Furthermore, through sensitivity analysis, we demonstrate its robustness to changes in hyperparameters compared to existing algorithms.
3. Several managerial insights into the operation of SSE are provided through the results of computational experiments. From the port's perspective, these insights help not only with the efficient operation of SSE, but also with the design and economic feasibility assessment for SSE and microgrids. The experimental results also provide insights for governments and institutions on implementing regulations or incentive policies.

The rest of this paper is structured as follows: In Section 2, we introduce existing studies on SSE and energy management. Our problem description and mathematical formulation for the problem are presented in Section 3. In Section 4, RL algorithms and a reward shaping technique for solving the MDP problem are proposed. Numerical experiments for the proposed algorithms are presented in Section 5, and Section 6 provides conclusions.

2. Literature review

This paper is related to two major research areas: decision-making in SSE operations and the energy management problem. In Section 2.1, we present existing studies that address various issues related to SSE operations, and in Section 2.2, we introduce research on energy management problems across different domains.

2.1. Decision-making in SSE operations

With growing environmental concerns and increasingly strict regulations in the maritime industry, significant studies have focused on the operation of SSE in ports. One of the widely discussed areas of research is the integration of SSE operations with other port operational problems. Zhen et al. [4,6] analyzed incentive policies aimed at promoting SSE, incorporating berth allocation and ship scheduling while considering environmental benefits and operational costs. Yu et al. [13,14] addressed the berth allocation and quay crane assignment problem with SSE operations, aiming to reduce emissions while satisfying the economic interests of both vessels and ports. They considered fluctuations in electricity prices to evaluate the economic benefits of SSE usage. These integrated studies focused on traditional port operational issues and did not take into account energy management for operating SSE.

Several studies have addressed energy management strategies for operating SSE through the port microgrids concept. Zhang et al. [15] focused on the synergy between SSE technology and microgrids, addressing both energy management and berth allocation to minimize operational costs. Wang et al. [11] addressed the design and operation of port microgrids and SSE, considering initial and operational costs. Using a two-stage optimization approach, they dealt with strategic decisions related to initial installation and operational level energy management. Additionally, several other studies have examined integrated energy management, including SSE, in conjunction with traditional port operations such as berth allocation [16] and crane assignment [17]. All of the above studies proposed ways to reduce operational costs by integrating port microgrids with SSE but did not account for SSE pricing to maximize profit. In contrast, Qiu et al. [18] focused on maximizing revenue from SSE supply for all-electric ships by determining pricing and electricity purchases from the main grid. However, the uncertainty of electricity prices in the real-time market was not considered in their study.

Additionally, there are studies that have explored the economic feasibility of adopting SSE. Yiğit et al. [19] assessed the economic feasibility of SSE from the perspective of shipping companies based on electricity and fuel prices, suggesting that SSE can provide economic benefits. Dai et al. [20] evaluated SSE from the port's perspective, considering carbon emission trading, and concluded that radical investments can yield economic gains with carbon emissions reduction. Case studies that conduct an economic assessment from the perspectives of both shipping companies and ports have also been presented [21]. Wang et al. [22,23] explored government SSE subsidy policies by analyzing the decision-making processes of governments, ports, and shipping companies using game theory. Xing et al. [24] also addressed the problem of ports deciding the production level and pricing of SSE, modeling it as a price-setting newsvendor problem that considers the uncertainty of conventional fuel prices. As an extension of these studies, Peng et al. [25] proposed a model in which the government determines subsidy policies, while the private/public port decides on investment scale and SSE pricing. While the aforementioned studies accounted for the profitability of SSE pricing for ports, they were conducted at a strategic level without considering the volatility of electricity prices.

2.2. Energy management problem

The majority of research related to energy management has focused on balancing energy supply and demand to reduce costs or increase profits. In microgrids, because the generation of renewable energy is uncontrollable, supply is regulated by deciding how much electricity to purchase from the main grid. At the same time, internal energy demand is managed directly, while external demand is adjusted indirectly through electricity pricing decisions. Accordingly, proper decision-making on both purchasing and pricing is essential for efficient energy management in microgrids. Joint decisions on dynamic pricing and purchasing have been widely studied in traditional inventory management in multi-echelon supply chain settings [26–29]. However, in contrast to these studies, the electricity market requires more dynamic and immediate decision-making due to the real-time fluctuations in wholesale prices.

Several studies on port microgrids have focused on enhancing energy efficiency through optimal system design and operation. Kumar et al. [30] addressed the design problem for port microgrids related to battery charging and SSE. Molavi et al. [12] proposed a two-stage programming approach to evaluate the practical viability of integrating microgrids into smart ports, considering both investment and operational decisions. More recently, there has been growing interest in optimizing the control strategies of distributed energy resources within port multi-energy systems to improve energy efficiency and reduce operational costs [31,32]. Although these studies dealt with operational issues related to energy management in port microgrids, they did not consider electricity trading with the main grid and thus did not cover pricing and purchasing decisions. Since there are few studies addressing the pricing and energy management problem for operating SSE in ports, we introduce studies that explore energy management strategies in other domains.

One of the main streams is the study of pricing strategies to control electricity demand in smart grids. Srinivasan et al. [33] investigated a pricing model for managing residential and commercial electricity demand within smart grids. A game theory-based methodology was proposed and the experiments demonstrated that real-time pricing policies can reduce peak loads while increasing the grid operator's profit. Lu et al. [34] explored a situation in which a smart grid operator determines retail electricity prices in a hierarchical electricity market, taking into account uncertainties in demand and wholesale electricity prices. This study aimed to optimize the overall system's profit by considering both the costs of the grid and customers. Zhang et al. [35] proposed an RL approach for dynamic pricing to maximize the profit of the broker acting as a retailer in the electricity market. Alves et al. [36] also addressed the electricity pricing problem for maximizing retailers' profit and developed a bilevel programming approach to model decision-making of both retailers and consumers. Electric vehicle (EV) charging stations, which purchase electricity from the main grid and resell it to EV users, have also been the subject of research on electricity pricing. Dong et al. [37] presented a simulation and optimization model for voltage control in EV charging stations through dynamic pricing. Several studies have also explored RL methodologies to find optimal pricing policies, taking into account various factors such as the profit of charging stations, the quality of service for EV users, and the utilization of charging facilities [38,39]. The aforementioned studies focused on pricing decisions to regulate demand and maximize profits; however, they did not address purchasing decision-making, as they considered situations without ESS.

Xu et al. [40,41] examined scenarios where retailers not only determine the selling price of electricity but also set the bidding price, which affects purchasing actions from the main grid. They developed a reinforcement learning model to simultaneously determine the bidding and selling prices for profit maximization. As noted in their research, in real-time electricity markets, a retailer's bidding can influence the wholesale price. However, in our study, we assume purchasing actions rather than bidding actions, because the amount of electricity purchased for SSE is not significant enough to impact the wholesale electricity price. In other words, the port is set to only take information on the wholesale price from the main grid, so bidding actions are out of our scope. Luo et al. [42] and Lee et al. [43] studied decision-making of EV charging stations similar to our study. They considered both electricity purchasing and pricing actions to maximize the profits of EV charging stations, while also taking renewable energy and ESS into consideration. However, the former study modeled real-time electricity prices using a Markov chain and assumed that the transition probability was known, while the latter assumed time-of-use (TOU) wholesale prices with lower uncertainty compared to real-time prices. To fill these research gaps, we address purchasing and pricing decisions under the uncertainty of the real-time electricity market without the unrealistic assumption on wholesale price.

3. Problem description and mathematical formulation

In this study, we consider an infinite-horizon dynamic pricing and energy management problem in a port microgrid aimed at maximizing the port's profit. Section 3.1 describes the dynamics of the system, including the operational structure of the port microgrid and the decision-making process of the vessels with respect to the use of SSE. Based on this, in Section 3.2, we formulate the problem as a Markov Decision Process (MDP) to enable a model-free approach. A list of notations used throughout the paper is provided in Table 1.

3.1. System dynamics

The SSE operation model involves three main entities: the main grid, the port, and the vessels. Electricity flows from the main grid through the port microgrid to the vessels, and the system is represented in a hierarchical structure as shown in Fig. 1. Based on the assumption mentioned in Section 2.2, the main grid sets the wholesale electricity price, c_t, based on hourly real-time pricing, and the port decides how much electricity to purchase, e_t, at that price. At the same time, the port decides the SSE price, p_t, and resells electricity to vessels. Consequently, the model can be shown as a multi-echelon supply chain, with the main grid acting as the wholesaler, the port as the retailer, and the vessels as the customers. In summary, at each time step t, the port makes two decisions to maximize the profit from operating SSE: (1) the amount of electricity to be purchased from the main grid, e_t, and (2) the selling price of SSE to vessels, p_t.

The port not only purchases electricity from the main grid but also generates it from its own renewable energy sources, such as solar and wind power. Accordingly, the port secures a renewable energy generation amount $r_t + \epsilon$ at each time step and stores it in an ESS. With the ESS, the port can purchase a large amount of electricity when the wholesale price is low and store the excess electricity for future sale. Typically, the electricity from the main grid or renewable energy sources, as well as the

Table 1
Notations.

Sets		
T	–	Set of time periods, $t \in T$
S_t	–	Set of states at time t
A_t	–	Set of actions at time t
Parameters		
e^{max}	–	Maximum amount of electricity to be purchased per unit time
d^{max}	–	Maximum amount of SSE to be sold per unit time
c_t	–	Real-time wholesale price of electricity at time t
I_t	–	Remaining electricity in ESS at the beginning of time t
I^{max}	–	Storage capacity of ESS
v_t	–	Number of vessels berthed at time t
D_t	–	Total electricity consumption of vessels at time t
r_t	–	Estimated renewable energy generation at time t
ϵ	–	Estimation error of renewable energy generation
$f_{n,t}$	–	Fuel price of vessel n at time t
η^{ch}	–	Charging efficiency of ESS (from AC to DC)
η^{disch}	–	Discharging efficiency of ESS (from DC to AC)
δ	–	Storage efficiency of ESS
$k_{n,t}$	–	Electricity consumption of a vessel n per unit time at time t
Decision variables		
e_t	Continuous	Amount of electricity to be purchased at time t
p_t	Continuous	SSE price at time t
d_t	Continuous	Amount of SSE sold at time t
$\lambda_{n,t}$	Binary	Whether vessel n is willing to use SSE at time t; 1 if $p_t \leq f_{n,t}$, 0 otherwise
π_t	Continuous	Profit of the port at time t

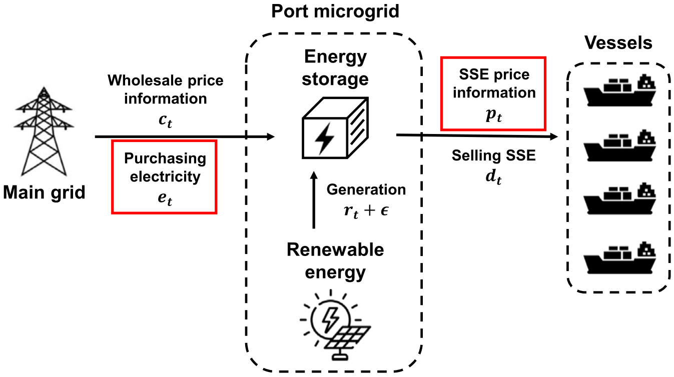


Fig. 1. Hierarchical structure of the pricing and energy management problem.

electricity consumed by vessels, is in AC form, whereas the ESS stores energy in DC form. Thus, power conversion is required to change AC to DC and then DC back to AC, resulting in electricity losses [44]. We account for these losses as the charging/discharging efficiency rate of the ESS, $\eta^{ch/disch}$ [16,18,42,45]. There are also electricity losses while the energy is stored in the ESS, which we consider as the storage efficiency rate, δ.

For each time step t, berthed vessels decide whether or not to use SSE based on the SSE price provided by the port. Each vessel compares its own fuel price $f_{n,t}$ with the SSE price p_t and chooses the more economical option. In reality, fuel prices can vary even for the same type of fuel depending on the port where the vessel is bunkered, and vessels may also obtain fuel at a specific price through contracts with their suppliers [46]. If $f_{n,t}$ is lower than p_t, the vessel n will generate electricity using its engines rather than using SSE. On the other hand, if $f_{n,t}$ is higher, the vessel will choose to use SSE. In short, each vessel's fuel price can be considered as the reservation price (willingness to pay) for using SSE, and we assume it follows a uniform distribution. We also assume that the port does not know the fuel price information for each vessel but only knows the probability distribution of the fuel prices. It is a realistic assumption that reflects the fact that shipping companies are reluctant to disclose information about their fuel prices. Under this assumption, the port does not make individual decisions for each vessel but instead considers the aggregated electricity demand from all berthed vessels when making pricing and purchasing decisions. As a result, the model avoids the combinatorial complexity that could arise when planning at the individual vessel level, especially when the number of vessels increases.

Based on the operational setting described above, the port's profit at time t is formulated as follows:

$$\pi_t = p_t d_t - c_t e_t, \qquad \forall t \in T, \tag{1}$$

where d_t denotes the amount of SSE sold to vessels at time t. The first term represents the revenue from SSE operations, and the second term denotes the cost incurred from purchasing electricity from the main grid.

The electricity flow constraints, including those related to the ESS, are defined as follows:

$$I_{t+1} \leq \delta \left(I_t + \eta^{ch}(e_t + r_t + \epsilon) - d_t/\eta^{disch} \right), \qquad \forall t \in T, \tag{2}$$

$$0 \leq I_t \leq I^{max}, \qquad \forall t \in T. \tag{3}$$

$$0 \leq e_t \leq e^{max}, \qquad \forall t \in T, \tag{4}$$

$$0 \leq d_t \leq d^{max}, \qquad \forall t \in T. \tag{5}$$

Constraint (2) is relevant to the ESS balance equation, which accounts for the inflows $(e_t + r_t + \epsilon)$ and outflows (d_t), adjusted by the charging, discharging, and storage efficiency factors. It is formulated as an inequality because the uncertain factors can cause the right-hand side to exceed the ESS capacity. Constraint (3) represents the storage capacity limit of the ESS. Constraints (4) and (5) specify the upper limits on the amount of electricity that can be purchased from the main grid and sold from the ESS to vessels, respectively.

The constraints related to the amount of SSE sold to vessels are as follows:

$$D_t = \sum_n k_{n,t}, \qquad \forall t \in T, \tag{6}$$

$$d_t \leq \sum_n k_{n,t} \lambda_{n,t}, \qquad \forall t \in T, \tag{7}$$

$$d_t \leq \gamma^{disch} I_t + \gamma^{ch} \gamma^{disch}(e_t + r_t + \epsilon), \qquad \forall t \in T, \tag{8}$$

$$0 \leq d_t \leq d^{max}, \qquad \forall t \in T. \tag{9}$$

In Eq. (6), D_t is defined as the total electricity consumption from the vessels at time t. Constraint (7) ensures that the amount of SSE sold does not exceed the total SSE demand from vessels. In this constraint, $k_{n,t}$ represents the electricity consumption of vessel n, and $\lambda_{n,t}$ is a binary variable indicating whether the vessel is willing to purchase SSE, determined by comparing p_t and $f_{n,t}$. Constraint (8) requires that the ESS discharging amount be less than or equal to the sum of the charged amount and the current storage level. Constraint (9) defines the maximum amount of SSE that can be sold.

To maximize the profit, the port makes joint decisions hourly on both e_t and p_t and each decision involves trade-offs. In the decision on e_t, purchasing a large amount when the wholesale price is low can reduce the purchasing cost. However, it leads to electricity losses due to the storage efficiency rate. In the decision on p_t, a lower price reduces the unit margin but increases sales volume, whereas a higher price increases the unit margin but reduces sales volume. Furthermore, when a large amount of electricity is stored in the ESS, it is expected to be more profitable to set a lower price to increase sales volume for reducing electricity losses. Conversely, if less electricity is stored, a higher price may be more beneficial. Accordingly, the port needs to consider various factors, including the wholesale electricity price, ESS storage level, renewable energy generation, and the electricity consumption of vessels.

3.2. Markov decision process

An MDP is a mathematical formalization of sequential decision making, which serves as the theoretical framework for reinforcement learning. MDPs consist of five key components: a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a state transition probability matrix $\mathcal{P}$, a reward function $\mathcal{R}$, and a discount factor γ. The state has the Markov property, which means the future state is independent of the past states given the current state. In other words, the state perfectly represents all the information about the current situation. Given the state, the agent interacts with the environment to maximize the sum of the cumulative rewards. Because the MDP problem forms the foundation of the reinforcement learning environment, developing it well is crucial for effectively solving the problem. In this section, we formulate the dynamic pricing and energy management problem as a Markov decision process (MDP) problem.

State: At time t, the state is defined as follows:

$$s_t = (c_{t-23}, c_{t-22}, \dots, c_{t-1}, c_t, I_t, r_t, D_t). \tag{10}$$

It contains information on the wholesale price, the ESS storage level, the estimated renewable energy generation, and the total electricity consumption. Specifically, we include not only the current wholesale electricity price but also its history over the past 23 h. Optimal decision-making requires information on future wholesale prices, but accurate prediction is challenging under real-time pricing. Instead, we use the wholesale price information from the most recent 24 h because real-time price data typically follows a daily pattern. Although the information on the fuel prices of the vessels is important for decision making, it is not included in the state because it is unobservable to the port.

Action: The action spaces are formulated as follows:

$$a_t = (e_t, p_t), \tag{11}$$

$$0 \le e_t \le e^{max}, \tag{12}$$

$$l \le p_t \le u. \tag{13}$$

Constraint (12) is the same as Constraint (4). As mentioned in Section 3.1, because the reservation prices of vessels for SSE follow a uniform distribution, there is a minimum value l and a maximum value u. Under this information, it is clear that the port would not set the SSE price lower than l or higher than u, so we set the feasible range of the SSE price as shown in Constraint (13). Several previous studies on energy management and pricing decisions use discrete action spaces for reinforcement learning. However, discrete action spaces have limitations when applied to real-world situations with a large action space. In this paper, we use continuous action spaces for both actions to ensure adaptability and applicability in practice.

Transition: In the MDP, the next state changes based on the action chosen by the agent. The transition probability refers to the likelihood that the next state will be s_{t+1} given that the agent takes action a_t in the current state s_t. In this problem, it is formulated as follows:

$$d_t = \min \left\{ d^{max},\ \sum_n k_{n,t}\lambda_{n,t},\ \gamma^{disch}(I_t + \gamma^{ch}(e_t + r_t + \epsilon)) \right\}, \tag{14}$$

$$I_{t+1} = \min\{I^{max},\ \delta(I_t + \gamma^{ch}(e_t + r_t + \epsilon) - (1/\gamma^{disch})d_t)\}. \tag{15}$$

Eq. (14) is defined by Constraints (7) through (9). Eq. (15) is derived from Constraints (2) and (3). Note that uncertainty of the environment arises not only from future information such as c_{t+1} and D_{t+1}, but also from the estimation error of renewable energy generation, ϵ, and unobservable factors, $\lambda_{n,t}$.

Reward: In the MDP, the agent receives a reward by taking action a_t in state s_t. The objective of this problem is to maximize the port's profit, so it is formulated as shown in Eq. (16), which is identical to Eq. (1).

$$r_t = \pi_t = p_t d_t - c_t e_t \tag{16}$$

4. Solution methods

In this section, we describe RL approaches to address the pricing and energy management problem for the operation of SSE in a port microgrid. When perfect information about the MDP is available, model-based algorithms such as dynamic programming can be utilized. However, uncertainties in the environment, such as wholesale electricity prices, renewable energy generation, and the total electricity consumption of vessels, lead to incomplete information about the MDP. Accordingly, we propose model-free RL algorithms that can be applied when the MDP is unknown.

RL is a type of machine learning in which an agent learns to maximize rewards by interacting with its environment. The agent selects an action from a set of feasible actions in the current state, receives a reward from the environment, and improves its policy. Through this process, the agent learns by trial and error, ultimately optimizing a policy for long-term rewards. Recently, with the advancement of deep learning, there has been significant research on deep RL (DRL) methods, which combine RL with deep learning techniques. In DRL, deep neural networks are used as function approximators for the actual value function or policy function. DRL methods can handle high-dimensional state spaces in real-world problems and are applicable even when the state spaces are continuous.

4.1. Actor-critic approach

As mentioned in Section 3.2, this problem involves a continuous action space. Value-based algorithms estimate the value function and select the action with the highest value for a given state. However, they are challenging to use when dealing with continuous action spaces, because the process of selecting the action with the highest value becomes an optimization problem. Another approach is to discretize the action space. However, this makes exploration and effective learning difficult, because it can significantly enlarge the action space due to the curse of dimensionality. Therefore, we adopt the actor-critic approach, which approximates both the value function and the policy function using neural networks, making it possible to handle continuous action spaces.

Most model-free RL algorithms iterate two key processes: policy evaluation, where the value of an action is estimated, and policy improvement, where the policy is updated based on the action-value function. The actor-critic approach consists of two networks: an actor network, which maps a state to a specific action, and a critic network, which estimates the value of the state-action pair. In this context, updating the critic network to obtain more accurate state-action values is referred to as policy evaluation, while updating the actor network based on those values is known as policy improvement. By iteratively updating both the actor and critic networks, the algorithm learns accurate value estimates and an optimal policy. Because the actor-critic approach trains both the actor and critic concurrently, it tends to be more stable than value-based or policy-based methods. Additionally, it has the advantage of being applicable even when the state and action spaces are large or continuous.

In this study, all algorithms use the same architecture for the actor and critic networks, as illustrated in Fig. 2. Each network consists of three hidden layers with 256 nodes, and the rectified linear unit (ReLU) function is used as the activation function for all hidden layers. The actor network takes the state as input and outputs the corresponding action, using the hyperbolic tangent (tanh) function as the output function. The critic network takes a state-action pair as input and outputs a Q-value. This architecture was chosen through repeated experiments to ensure training stability.

4.2. Deep deterministic policy gradient algorithm

Recent RL research predominantly focuses on DRL methods, with the first algorithm to successfully implement DRL being the deep Q-network (DQN). DQN is an innovative algorithm capable of addressing

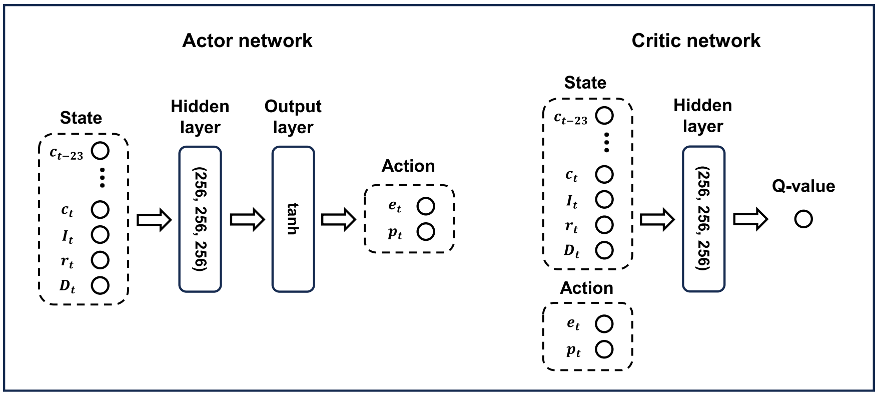


Fig. 2. Architecture of the actor and critic networks.

problems with large or continuous state spaces. However, DQN remains difficult to apply to continuous action spaces because it is a value-based algorithm. Lillicrap et al. [47] proposed the deep deterministic policy gradient (DDPG) algorithm to overcome this limitation by combining the actor-critic approach with two key techniques from DQN. One of these techniques is experience replay, which enhances the efficiency of learning. The agent interacts with the environment to gain experiences during the learning process, and these experiences are stored in a replay buffer to be reused in training. It not only enhances data efficiency but also reduces the correlation between experiences by sampling randomly from the buffer, thereby increasing the overall efficiency of learning. The other technique is use of a target network during the network training process. The actor-critic approach involves training both the actor network π_ϕ and the critic network Q_θ, which are associated with their own target networks $\pi_{\phi'}$, and $Q_{\theta'}$. If the target values continuously change during the network training process, it can lead to instability in learning. The target network is updated at fixed intervals rather than continuously, which improves the stability of learning.

In the DDPG algorithm, the temporal difference (TD) target is used to train the critic network. The TD target is an estimate of the current step's value, based on the actual reward from the current step and the estimated value of the next step. The value of the next step is estimated through the critic target network, and the TD target can ultimately be expressed as $y_t = r_t + \gamma Q_{\theta'}(s_{t+1}, \pi_{\phi'}(s_{t+1}))$. Therefore, the critic loss function can be derived as shown in Eq. (17), and the critic network is updated using a gradient descent, as shown in Eq. (18).

$$L(\theta) = \mathbb{E}_\pi[(y - Q_\theta(s,a))^2] = N^{-1}\sum[(r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s')) - Q_\theta(s,a))^2] \tag{17}$$

$$\theta \leftarrow \theta + \alpha \sum[(r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s')) - Q_\theta(s,a))\nabla_\theta Q_\theta(s,a)] \tag{18}$$

In problems involving discrete action spaces, policy improvement is achieved by solving for the action with the highest value in a given state, $argmax_a Q^\pi(s,a)$. However, in continuous action spaces, finding the optimal action becomes an additional optimization problem. Silver et al. [48] proposed an alternative method where the policy is updated in the direction of the gradient of the value function Q, and derived Eq. (19), the gradient of the policy's value function $J(\phi)$, using the deterministic policy gradient (DPG) theorem. This outlines the theoretical background of the DDPG algorithm, where the actor network is updated based on the gradient ascent method, as shown in Eq. (20).

$$\nabla_\phi J(\phi) = N^{-1}\sum \nabla_a Q_{\theta_1}(s,a)\nabla_\phi \pi_\phi(s) \tag{19}$$

$$\phi \leftarrow \phi + \beta \nabla_\phi J(\phi) \tag{20}$$

Exploration is a critical and widely addressed issue in RL. RL agents learn through experiences, so they cannot obtain the value of states and actions they have never encountered. The DDPG algorithm enables exploration by adding random noise $\mathcal{G}$ to the actions generated by the actor network's output. The detailed procedure of the DDPG algorithm is presented in Algorithm 1.

4.3. Twin delayed DDPG algorithm

While numerous studies have employed the DDPG algorithm to address continuous control problems, challenges such as converging to local optimal solutions or divergence still remain. Fujimoto et al. [49] proved that these issues are due to the overestimation of Q-values in the DDPG algorithm and proposed the twin delayed DDPG (TD3) algorithm, which incorporates several techniques to address this limitation. The TD3 algorithm solves the overestimation bias problem by adopting the concept of double Q-learning proposed by Ref. [50] during the critic network update process. Specifically, it introduces two independent critic networks $Q_{\theta'_1}$, $Q_{\theta'_2}$, and when updating the target, the algorithm selects the minimum of the two estimated values. In addition, to prevent overfitting in the value estimation process, noise $\mathcal{G}$ is added to the original action. This technique helps ensure that similar actions yield similar value estimates in continuous action space environments. Accordingly, the update process used in the DDPG algorithm, as presented in Eq. (18), is replaced by that in Eq. (21).

Algorithm 1 DDPG algorithm.

```
Initialize actor network $\pi_\phi$, critic network $Q_\theta$ with random parameters $\phi$, $\theta$
Initialize actor target network $\pi_{\phi'}$, critic target network $Q_{\theta'}$ with parameters $\phi' \leftarrow \phi$, $\theta' \leftarrow \theta$.
Initialize replay buffer $\mathcal{B}$
for episode $e = 1$ to $E$ do
    Initialize electricity price, renewable energy generation, total electricity consumption of vessels, and ESS storage
    for time step $t = 1$ to $T$ do
        Observe state $s_t$
        Select action $a_t = \pi_\phi(s_t) + \mathcal{G}$ with exploration noise $\mathcal{G} \sim \mathcal{N}(0, \sigma)$.
        Execute action $a_t$ and get reward $r_t$, next state $s_{t+1}$.
        Add transition $(s_t, a_t, r_t, s_{t+1})$ to $\mathcal{B}$.
        Sample mini-batch $(s, a, r, s')$ from $\mathcal{B}$ with batch size $N$.
        Update critic network $\theta$ using Eq. (18).
        Update actor network $\phi$ using Eqs. (19) and (20).
        Update actor target network $\phi' = \tau\phi + (1 - \tau)\phi'$
        Update critic target networks $\theta' = \tau\theta + (1 - \tau)\theta'$
    end
end
```

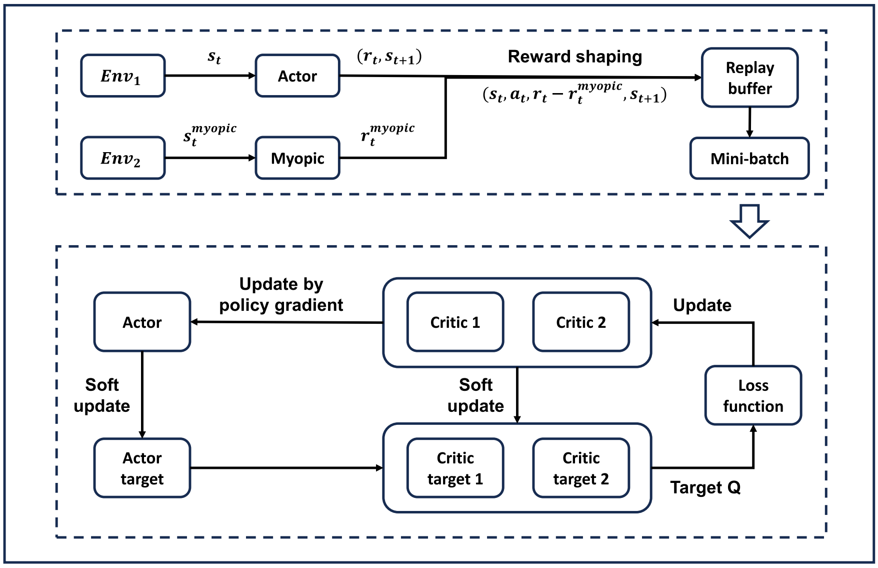


Fig. 3. Overview of the TD3 algorithm with reward shaping.

$$\theta_i \leftarrow \theta_i + \alpha \sum [(r + \gamma min_{i=1,2} Q_{\theta'_i}(s', \pi_{\phi'}(s') + \mathcal{G}) - Q_{\theta_i}(s,a)) \nabla_{\theta_i} Q_{\theta_i}(s,a)] \tag{21}$$

In the DDPG algorithm, both the critic and actor networks are updated at every time step. The actor network is updated based on the value estimates provided by the critic network, so if the critic network is inaccurate, the learning of the actor network also becomes unreliable. The TD3 algorithm addresses this issue by recognizing that such a process can lead to instability in learning and instead adopts a method where the actor network is updated after the critic network has been trained more stably. Specifically, the critic network is updated at every time step, while the actor network is updated at constant intervals.

4.4. Reward shaping technique

We propose a model that combines the TD3 algorithm with a reward shaping technique to achieve better performance. Reward shaping is a method used in RL to incorporate knowledge from an external policy when applying RL to a specific domain. Since its theoretical introduction by Ref. [51], this technique has been widely utilized in various fields, such as inventory management [52,53] and energy management [54]. Reward shaping changes the original reward function R to a shaped reward function R' by adding a shaping function F. In this study, we use a myopic algorithm that maximizes the expected reward in the current step as a baseline policy, and the shaped reward function is defined as shown in Eq. (22). In Eq. (22), it is constructed by subtracting the profit obtained by myopic policy in time t from that obtained by TD3 policy in time t. We provide a detailed explanation of the myopic algorithm in Appendix A.

$$r'_t = \pi_t - \pi_t^{myopic} \tag{22}$$

This shaped reward can be interpreted as the relative reward of the TD3 policy compared to the myopic policy. In this study, the myopic policy used for reward shaping is executed in a separate environment, making it independent of the actions selected by the TD3 policy. Using such a shaped reward function offers two key advantages for the model. First, in line with the purpose of reward shaping mentioned earlier, our model learns a better policy by leveraging the knowledge from the myopic policy. We expect the myopic policy's reward to serve as a minimum performance baseline for policy learning. Second, more importantly, the shaped reward function offers a more accurate representation of the action's value.

As mentioned in Section 4.1, the key principle of the actor-critic approach is to find an action that maximizes the expected cumulative reward for a given state. However, in this problem, the reward is influenced by external factors such as the wholesale electricity price and renewable energy generation, rather than the agent's actions. Motivated by the above observation, we designed the reward shaping technique to remove the impact of external factors from the actual reward. In this context, we use the myopic algorithm as a baseline because it provides the most direct estimate of the current state's value. This idea is similar to the concept of advantage actor-critic (A2C) algorithms proposed by Ref. [55], where the advantage is calculated by subtracting the state value from the state-action value. While the A2C algorithm requires training an additional neural network to estimate the state value, the reward shaping technique proposed in this paper enables stable estimation of the state value using a myopic algorithm. The overview of the TD3 algorithm with reward shaping is shown in Fig. 3, and the detailed learning procedure is presented in Algorithm 2.

5. Computational experiments

In this section, we conducted computational experiments to evaluate and analyze the performance of the proposed RL algorithms, DDPG, TD3, and TD3 with reward shaping (TD3-RS). The experiments consisted of three main parts. In Section 5.1, we compared the performance of the three proposed RL algorithms and validated the algorithms using a test dataset. In Section 5.2, we performed a sensitivity analysis on several environmental inputs. Finally, in Section 5.3, we analyzed the potential side effects that arise from the port's pricing and energy management strategies.

We established the following common experimental settings across the three experiments. Based on the problem description outlined in Section 3.1, we implemented an environment for the dynamic pricing and energy management problem. The electricity wholesale price data used in the environment were obtained from the website of PJM, one of the regional transmission organizations in the United States, and real-time hourly electricity pricing data were used. The renewable energy generation data were also generated based on PJM's hourly renewable energy generation dataset, while the number of berthed vessels was derived from data on container ships berthed at the Port of Los Angeles. The parameters for the upper and lower bounds of vessel fuel prices were

Algorithm 2 TD3 algorithm with reward shaping.

```
Initialize actor network $\pi_\phi$, critic networks $Q_{\theta_1}$, $Q_{\theta_2}$ with random parameters $\phi$, $\theta_1$, $\theta_2$
Initialize actor target network $\pi_{\phi'}$, critic target networks $Q_{\theta'_1}$, $Q_{\theta'_2}$ with parameters $\phi' \leftarrow \phi$, $\theta'_1 \leftarrow \theta_1$, $\theta'_2 \leftarrow \theta_2$.
Initialize replay buffer $\mathcal{B}$
Initialize policy update frequency $d$
for episode $e = 1$ to $E$ do
    Initialize electricity price, renewable energy generation, total electricity consumption of vessels, and ESS storage for two environments: $Env$, $Env_2$
    for time step $t = 1$ to $T$ do
        Observe state $s_t$ of $Env_1$ and select action $a_t = \pi_\phi(s_t) + \mathcal{G}$ with exploration noise $\mathcal{G} \sim \mathcal{N}(0, \sigma)$.
        Execute action $a_t$ in $Env_1$ and get reward $r_t$, next state $s_{t+1}$.
        Observe state $s_t^{myopic}$ of $Env_2$ and select action $a_t^{myopic}$ using the myopic algorithm.
        Execute action $a_t^{myopic}$ in $Env_2$ and receive reward $r_t^{myopic}$
        Shaped reward $r'_t = r_t - r_t^{myopic}$
        Add transition $(s_t, a_t, r'_t, s_{t+1})$ to $\mathcal{B}$.
        Sample mini-batch $(s, a, r, s')$ from $\mathcal{B}$ with batch size $N$.
        Update critic networks by Eq. (21) with policy noise $\tilde{\mathcal{G}} \sim clip(\mathcal{N}(0, \tilde{\sigma}), -w, w)$
        if $t \bmod d = 0$ then
            Update actor network by policy gradient: $\nabla_\phi J(\phi) = N^{-1} \sum \nabla_a Q_{\theta_1}(s, a) \nabla_\phi \pi_\phi(s)$
            Update actor target network $\phi' = \tau\phi + (1-\tau)\phi'$
            Update critic target networks $\theta'_i = \tau\theta_i + (1-\tau)\theta'_i$
        end
    end
end
```

estimated using data from the website of Oilmonster, which provides information on regional bunker prices. Parameters related to the ESS were set based on data from the website of Hyosung Heavy Industries, an energy solution company in South Korea. Most experiments were conducted using an AMD Ryzen 5 7600X 6-Core Processor with Python version 3.12, and the myopic algorithm was solved with FICO Xpress version 8.14.

5.1. Performance evaluation of the proposed RL algorithms

In this subsection, we compared the performance of the three RL algorithms with that of the time-segmented heuristic algorithm described in Appendix B. Although we assumed that vessel fuel prices follow a uniform distribution, we additionally conducted experiments under the assumption that the prices follow a truncated normal distribution, aiming to demonstrate the robustness of the algorithm to the underlying price distribution. Because the cumulative distribution function of the truncated normal distribution does not have a closed-form expression, we used the Abramowitz & Stegun approximation [56] to represent Constraint (A.2) in the myopic algorithm. In these experiments, each episode had a length of 50 days, corresponding to 1200 time steps, and the model was trained over 5000 episodes. During the training process, the performance of the model was evaluated in a separate evaluation environment every 10 episodes, and this performance was used as a metric for evaluation. We determined hyperparameters for the RL algorithms through experiments as provided in Table 2, and these were identically applied to all three algorithms.

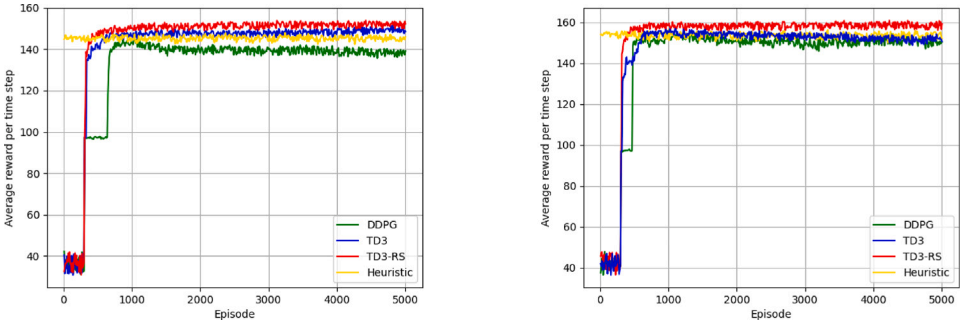
Fig. 4 (a) shows the learning curves of the three RL algorithms and the result of the heuristic algorithm, under the assumption that vessel fuel prices follow a uniform distribution. The X-axis represents the number of episodes during the training, while the Y-axis indicates the average reward per time step. The DDPG algorithm, depicted by the green graph, showed the lowest performance, even below that of the heuristic algorithm. Although it achieved a high reward around the 1000th episode, the reward decreased as training progressed and finally converged to a lower reward, which indicated instability in the learning process. With the TD3 algorithm, represented by the blue graph, the learning process was relatively stable, and it converged after 4000 episodes. The average reward after convergence was $149.08, slightly higher than the average reward of the heuristic algorithm, which was $145.26. The result of the TD3-RS is shown by the red graph, and it converged quickly and stably within 2000 episodes. It achieved the highest performance and the average reward after convergence was $151.69, approximately $6.43 higher than that of the heuristic algorithm. Even though this difference may seem to be trivial, considering that this value represents the reward per time step (one h), it becomes a significant amount when extended to a day and a year. Fig. 4 (b) presents the results obtained under the truncated normal distribution assumption. Consistent with the previous experiment, TD3-RS algorithm demonstrated the best performance among the evaluated algorithms. The computational times of training 5000 episodes for DDPG and TD3 algorithms were approximately 10 h and 11 h, respectively. Although there may have been concerns about the computational time required for the TD3-RS algorithm due to the process of solving NLP problems, the problem solved by the myopic algorithm was a small-size problem that could be solved very quickly using a solver, resulting in a total computational time of 13 h.

Table 2
Hyperparameters of the RL algorithms.

Hyperparameter	Value
Capacity of the experience replay buffer	1,000,000
Size of the sampled mini-batch	256
Discount factor γ	0.99
Soft update parameter τ	0.005
Learning rate of actor and critic networks	0.0001
Exploration noise parameter σ	0.2
Policy noise parameter $\tilde{\sigma}$	0.1
Policy noise parameter w	0.1
Frequency of policy updates d	2

We also conducted a sensitivity analysis of hyperparameters for the three RL algorithms. For each algorithm, the learning process was performed under different values of the standard deviation of the policy noise (0.1, 0.2, 0.3) and the capacity of the replay buffer (100,000, 500,000, 1000,000). Fig. 5 illustrates the average reward per time step after convergence for each experimental setting. The DDPG algorithm exhibited consistently lower performance than the other algorithms across all settings and was especially sensitive to changes in the replay buffer capacity. The TD3 algorithm achieved performance close to that of the TD3-RS algorithm in certain settings, specifically when the policy noise standard deviation was set to 0.2 and the replay buffer capacity to 10,00,000. However, it showed high sensitivity to both hyperparameters overall. In contrast, we observed that the TD3-RS algorithm maintained high performance across changes in both hyperparameters, indicating its robustness to hyperparameter settings.

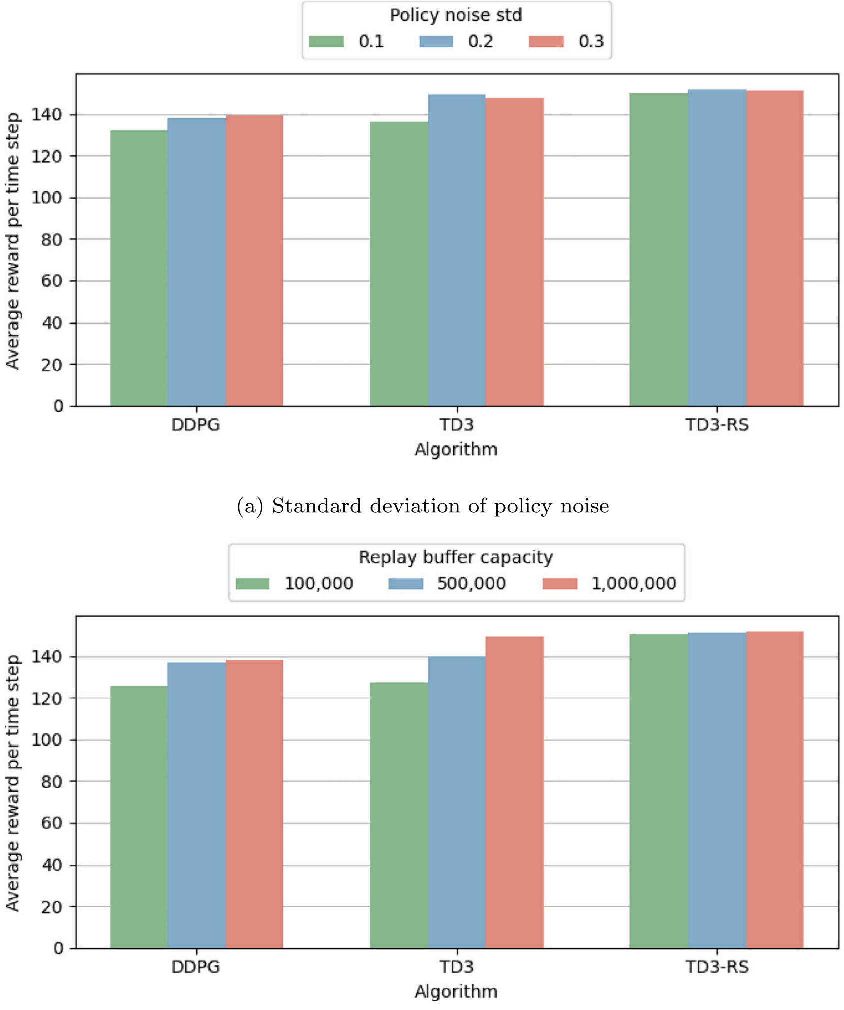
RL models trained with the training dataset may suffer from an overfitting problem, so we conducted validation experiments using a test dataset. We also verified the economic impact of the ESS through these experiments. While the ESS can benefit from fluctuations in wholesale prices, it also has drawbacks due to charging and discharging electricity losses. We executed experiments on a scenario where the port determines pricing and supplies SSE to vessels without using an ESS, and compared the results with those obtained from the RL algorithms. We classified the test dataset into four cases based on the mean and variance of wholesale electricity prices, and validated the algorithms for each case. The four cases were as follows: HH case, HL case, LH case, and LL case, respectively representing high mean and high variance; high mean and low variance; low mean and high variance; and low mean and



(a) Uniform distribution

(b) Truncated normal distribution

Fig. 4. Performance comparison of algorithms during the training process.



(a) Standard deviation of policy noise

(b) Capacity of replay buffer

Fig. 5. Sensitivity analysis of hyperparameters.

low variance in wholesale electricity prices. The test data for each case had a length of 30 days (720 time steps), and the means and standard deviations of electricity prices within each case are specified in Table 3.

We repeated each validation experiment 10 times for all RL algorithms and cases. The average reward per time step and the difference from the results of the scenario without an ESS are presented in Table 4. The validation experiment results showed that, similar to the findings of the previous experiment, the TD3-RS algorithm achieved the highest performance in all cases. In particular, the results of the TD3-RS algorithm were considerably higher than those of the scenario without the ESS for all cases, implying the economic benefits of an ESS. In comparing the results of the TD3-RS algorithm across different cases, the algorithm demonstrated relatively better performance in cases with high variance in wholesale electricity prices (HH, LH cases) compared to those with

Table 3
Means and standard deviations of wholesale electricity prices.

Case	Mean	Std
HH case	42.38	38.93
HL case	41.96	19.26
LH case	27.29	27.57
LL case	28.44	13.51

low variance (HL, LL cases). These results suggest that the algorithm takes advantage of purchasing electricity at low prices and storing it for sale when prices are high. In HH and LH cases, the significant gap between high and low wholesale prices enables the port to take full advantage of these benefits, while such benefits are less pronounced in HL and LL cases.

Fig. 6 depicts the changes in wholesale electricity prices over time, while Fig. 7 illustrates the changes in ESS storage levels when using the TD3-RS algorithm. In Fig. 6, the wholesale prices in all four cases commonly exhibit a pattern of being low during early morning hours, increasing until evening, and then decreasing again. Due to this pattern, the storage levels in Fig. 7 tend to increase during early morning hours and decrease during the daytime when wholesale prices increase. These results align with our intuition regarding the ideal ESS storage levels. They also support the results of the validation experiments. In Fig. 6, the high variance cases, such as the HH case (red graph) and the LH case (green graph), show large differences between the peaks and valleys of wholesale prices. Conversely, the low variance cases, such as the HL case (blue graph) and the LL case (yellow graph), have smaller differences. As a result, in Fig. 7, ESS storage levels in the low variance cases are generally lower than those in the high variance cases.

Fig. 7 not only depicts the trends in ESS storage levels but also provides insights into the appropriate storage capacity for the ESS. The ESS is a facility with high initial costs and its storage capacity has a significant impact on overall costs. Thus, it is a critical decision factor when installing an ESS at a port. In this experiment, the ESS storage capacity was set to 50 MWh. However, if the storage levels resulting from the application of the algorithm are lower than this value, it is proper to select a lower initial storage capacity.

5.2. Sensitivity analysis

We conducted a sensitivity analysis on three factors: the electricity consumption of vessels, the distribution of vessel fuel prices, and the emission weight parameter. These factors either influence the demand for SSE or directly affect the reward function, thereby impacting the port's decision-making process. We trained the TD3-RS algorithm under varying values of each factor and compared the results. Each experiment illustrates how different operating conditions affect the port's decisions and overall profit.

We compared the average reward for different mean values of vessel's electricity consumption per unit time, specifically 0.75, 1.0, 1.25, and 1.5. The results are illustrated in Fig. 8. We observed that as the vessel's electricity consumption increased, the average reward also increased. This is reasonable, given that the increase in the overall use of electricity by vessels directly translates into increased SSE demand. In addition, the results showed that the marginal increase in reward gradually decreased. It could be interpreted as being due to constraints such as the ESS capacity and the upper limit on SSE sales.

In the sensitivity analysis on the standard deviation of vessel fuel prices, we conducted experiments using truncated normal distributions with standard deviations of 1, 3, and 5, as well as a uniform distribution. The uniform distribution has a larger standard deviation than the truncated normal distributions. As shown in Fig. 9, the port's profit decreased as the standard deviation of fuel prices increased. Because the port does not have information on each vessel's fuel price, higher variability increases environmental uncertainty, ultimately leading to a decrease in profit.

All of the above experiments were conducted with the objective of maximizing the port's economic profit. However, in reality, ports operate SSE not only for economic gains but also to achieve environmental benefits. Specifically, this trend has grown with the rise of carbon credit trading, and many studies related to supply chain issues have incorpo-

Table 4
Validation results: means and differences from the scenario without ESS.

Case	Without ESS	DDPG		TD3		TD3-RS	
		Mean	Difference	Mean	Difference	Mean	Difference
HH case	140.35	140.83	0.34 %	157.41	12.16 %	159.35	13.54 %
HL case	100.59	93.59	−6.96 %	107.57	6.94 %	109.12	8.48 %
LH case	193.08	197.61	2.35 %	209.66	8.59 %	212.56	10.09 %
LL case	148.87	144.42	−2.99 %	158.23	6.29 %	161.03	8.17 %

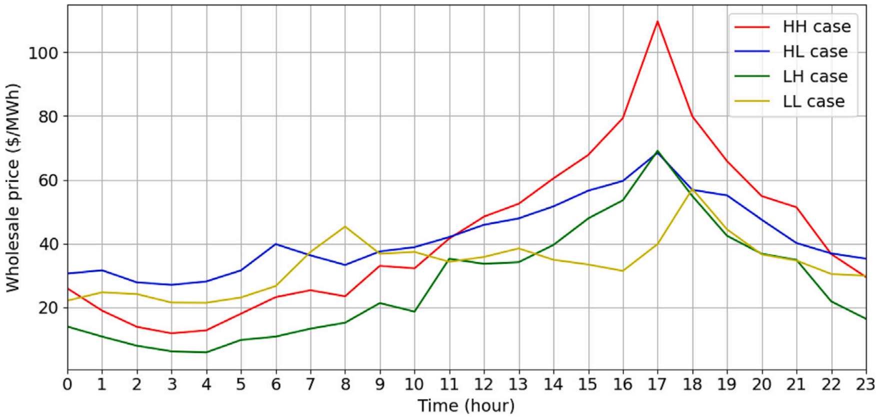


Fig. 6. Wholesale electricity price over time for each case.

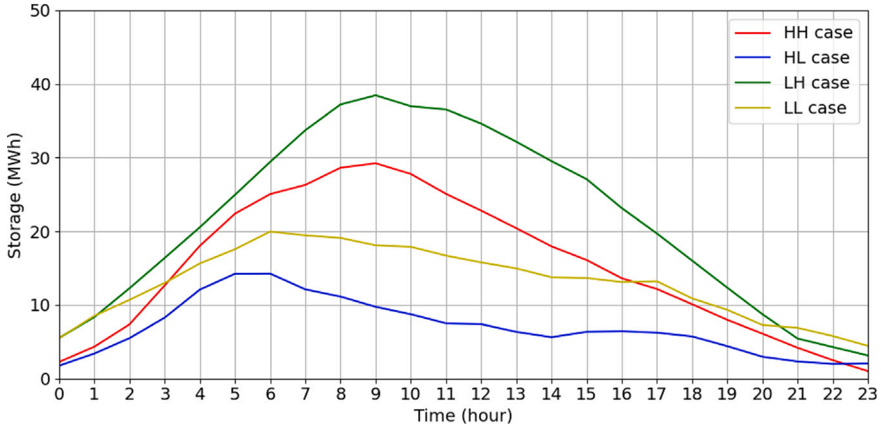


Fig. 7. ESS storage level over time for each case.

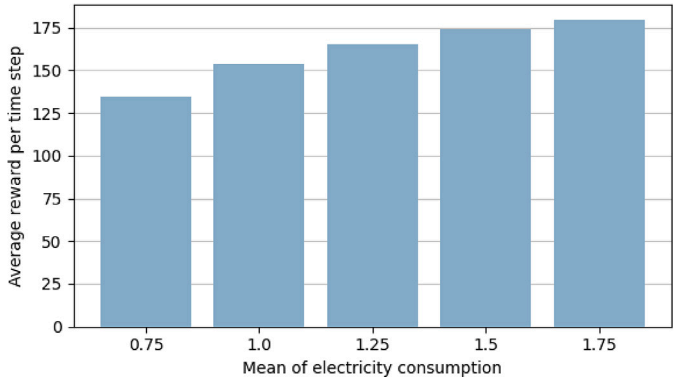


Fig. 8. Sensitivity analysis on mean of vessel electricity consumption.

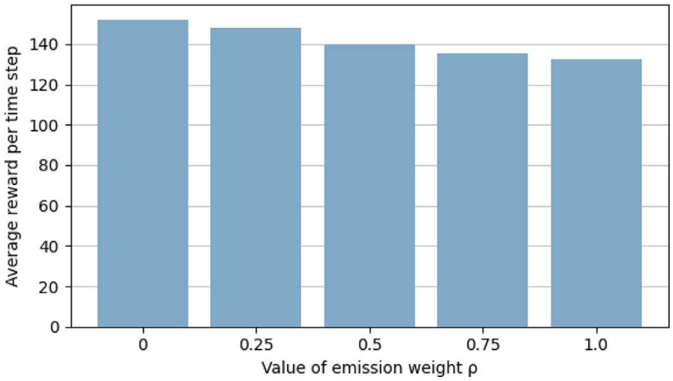


Fig. 10. Sensitivity analysis on emission weight ρ.

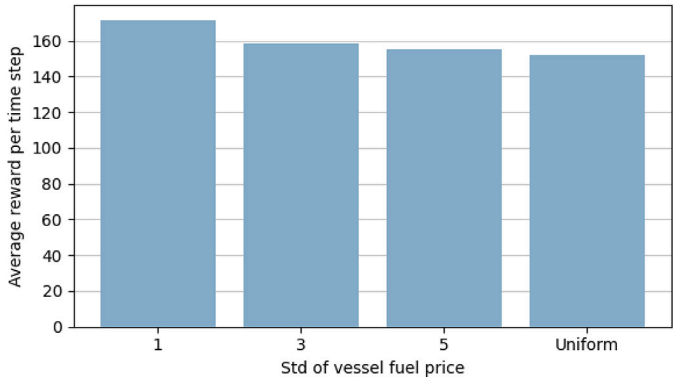


Fig. 9. Sensitivity analysis on standard deviation of vessel fuel price.

rated emission factors, such as carbon taxes and carbon credit trading [57,58]. In this section, we incorporated an emission factor into the existing reward function, using a revised reward function that reflects both economic and environmental benefits. The revised reward function is defined as follows:

$$r_t = \pi_t + \rho \cdot d_t \cdot benefit^{env} \tag{23}$$

where ρ represents the weight assigned to the port's environmental benefits, d_t is the sales volume of SSE, and $benefit^{env}$ means the social and environmental benefits per unit of SSE sold. Based on the previous study [21], we set $benefit^{env} = 22$ ($\$/MWh$). The value of ρ varies depending on the port's emphasis on environmental factors, as well as on additional benefits from SSE sales through carbon credit trading, the regulatory intensity imposed by the government, and other factors.

We carried out experiments for $\rho = 0, 0.25, 0.5, 0.75,$ and 1.0, and the results are presented in Fig. 10. It should be noted that the average reward shown in Fig. 10 does not include environmental benefits. The results showed that the port's profit decreased as ρ increased. This can be explained by the port's incentive to increase SSE sales, even at the cost of higher electricity purchases and storage levels.

Fig. 11 shows the average ESS storage level for each value of ρ, supporting the result of Fig. 10. Except for the interval where ρ increases from 0.75 to 1.0, the average ESS storage level increased as the value of ρ increased. This aligns with the expectation that a higher ρ would make the port store larger amounts in the ESS to sell SSE even when future wholesale prices are high. However, contrary to expectations, when ρ increased from 0.75 to 1.0, the average storage level remained almost the same. It was due to the limitations imposed by the storage capacity of the ESS and the upper bound on the amount of electricity that could be purchased from the main grid per unit time. If the port was already fully utilizing the ESS when ρ was 0.75, an increase in ρ beyond this value would not significantly affect the port's policy. If the ESS capacity and the electricity purchase limit had been set higher, the average storage level would have continued to increase. These results imply that a port's emission weight also should be considered when determining the initial design of the ESS.

5.3. Analysis of side effects

In this subsection, we analyzed the potential side effects of the port's decisions on dynamic pricing and energy management for operating SSE.

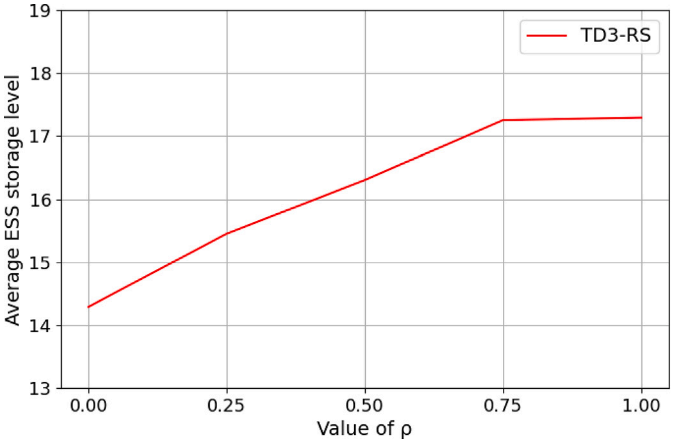


Fig. 11. Average ESS storage level with different emission weights.

While we have mainly focused on maximizing the port's profit, it is also important to consider how these decisions might impact stakeholders and what socio-environmental outcomes they might evoke. Ideally, the proposed algorithm helps with efficient energy management, so other stakeholders, such as shipping companies would benefit from it. We examined the side effects from three perspectives: the utilization of SSE, the costs incurred by vessels berthed at the port, and the profit of the entire system. Vessel costs refer to the sum of costs associated with using SSE and the fuel costs used for electricity generation. For analysis, we also conducted experiments on a scenario where the port did not decide the pricing for SSE, and vessels used SSE at the wholesale electricity price, which is referred to as a "scenario without pricing." The scenario without pricing is a common situation observed at many ports, and comparing our model with it clearly highlights the benefits of port microgrids. We compared the side effects of using the TD3-RS algorithm with those of the scenario without pricing.

Fig. 12 shows how the SSE utilization rate changes with varying values of ρ. Note that, in the scenario without pricing, the utilization rate remains constant regardless of changes in ρ because the port does not set prices. The results demonstrated that utilization rate when using the TD3-RS algorithm was higher than that of the scenario without pricing, even when $\rho = 0$. It indicated that the port could increase SSE utilization through the advantage of the ESS, even when making decisions only for its own economic benefit. This analysis of side effects can also provide insights for governments and institutions implementing regulations or incentive policies. For example, based on these results, if the government aims to increase the SSE utilization rate to 80 %, it would need to intensify regulations or provide incentives to raise the port's emission weight to above 0.5.

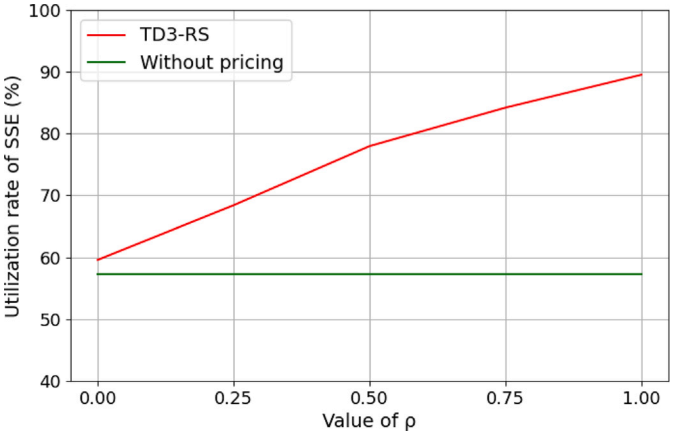


Fig. 12. Utilization rate of SSE with different emission weights.

Fig. 13 presents the results of the cost analysis for berthed vessels. As ρ increased, the costs for vessels decreased because the port lowered prices to increase the environmental benefits as ρ increased. However, these costs are substantially higher than the vessel costs in the scenario without pricing. This indicates that the port's pricing strategy leads to increased costs for vessels even when the emission weight is high due to policies from institutions or the government. On the other hand, Fig. 14 represents the profit of the entire system, calculated by subtracting the vessels' costs from the port's economic profit. The results show that the entire system profit is highest when using the TD3-RS algorithm and significantly higher compared to the scenario without pricing. This reveals that while the port's pricing strategy increases vessel costs, it is beneficial from the perspective of the entire system. Accordingly, if the port provides appropriate incentives for vessels to use SSE, the port's pricing and energy management strategy can benefit both the port and shipping companies. Moreover, the profit of the entire system increases as the emission weight rises from 0 to 0.25, but once it exceeds a certain level, it begins to decrease. This means that excessively strict regulations or incentives may reduce the overall system's profit, so it is important to consider this when establishing appropriate policies.

5.4. *Managerial insights*

Based on the results of the above experiments, we provide managerial insights not only for ports that are operating or planning to adopt SSE

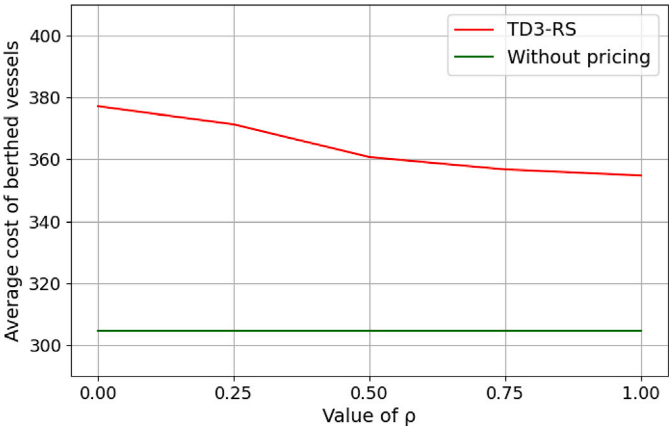


Fig. 13. Average cost of berthed vessels with different emission weights.

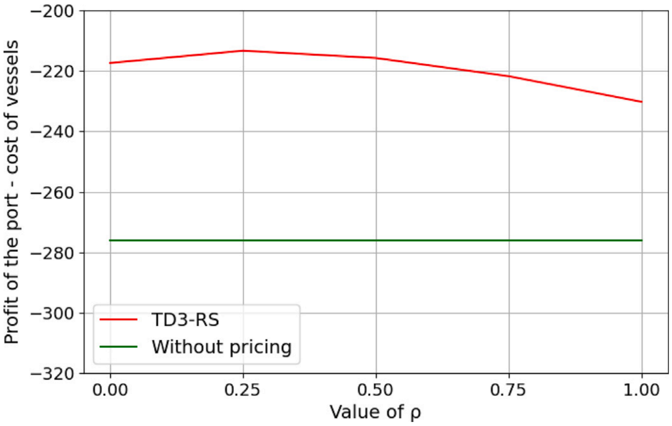


Fig. 14. Profit of the entire system with different emission weights.

but also for governments and institutions implementing regulatory or incentive policies.

1. The characteristics of electricity prices vary depending on the structure and policies of the national or regional electricity market. Through the experiments outlined in Section 5.1, we confirmed that the economic benefits of ESS are greater when there is high volatility in wholesale electricity prices. Therefore, considering the initial cost of ESS, we recommend that ports in regions with high electricity price volatility use ESS for operating SSE.
2. Determining the appropriate storage capacity when installing an ESS at a port reduces unnecessary costs and enables an efficient microgrid design. In Sections 5.1 and 5.2, we analyzed ESS storage levels based on the characteristics of wholesale electricity prices and emission factors, respectively. When price volatility is high and emission weight is large, the benefits of energy storage are substantial, making it advisable to choose greater storage capacity. Conversely, in cases of low price volatility and a smaller emission weight, an ESS with smaller capacity is recommended.
3. The experimental results in Section 5.3 show that while regulations or incentives for ports certainly increase the SSE utilization rate, they have limitations in reducing costs for vessels. Although ports could lower vessel costs by providing incentives for using SSE, institutions or governments cannot enforce it directly. A reasonable approach would be for institutions or governments to implement appropriate regulations on ports to maximize the profit of the entire system, while introducing incentive policies for vessels to use SSE. This would result in greater benefits for both vessels and ports compared to the scenario without pricing.

6. Conclusions

As the maritime industry moves toward reducing emissions, SSE technology has emerged as a key solution for port decarbonization. Although governments and organizations worldwide are promoting SSE adoption through regulations and incentive policies, the utilization rate remains low due to high initial and operational costs. To solve this fundamental problem, a profitable model for SSE operation is required, and we focused on the economic benefits generated by integrating SSE with port microgrids. In this study, we proposed a dynamic pricing and energy management strategy to maximize the profitability of the port's SSE operation. To the best of our knowledge, this is the first research to address dynamic pricing and energy management from a retailer's perspective in a real-time electricity market without the unrealistic assumption of predicting electricity prices. To address this challenging problem, we adopted an actor-critic RL approach and applied reward shaping to improve the model's performance and stability.

We used three RL algorithms, DDPG, TD3, and TD3-RS, and demonstrated through computational experiments that our proposed TD3-RS algorithm outperforms the others. In particular, we evaluated the algorithms' performance in different cases based on the characteristics of wholesale electricity prices. The results showed that the utility of an ESS is higher when electricity prices are low and volatile, while the advantage of using an ESS is limited in the opposite case. The proposed algorithm also exhibited strong performance with a revised reward function that incorporates the emission factor, demonstrating its robustness. Additional experiments suggest that strategies for maximizing a port's profits have the potential to increase the utilization of SSE and enhance the overall benefits of the system, leading to positive side effects.

The primary cause of the currently low SSE utilization rate is the low installation rate of SSE systems in ports and vessels. The installation rate at ports could be increased by proposing a profitable model and algorithms for operating SSE. Additionally, the findings of this study provide insights for regulatory and incentive policy decisions related to SSE adoption. At the regional level, public benefits such as higher SSE utilization or higher overall system profits can be achieved by properly adjusting the intensity of policies according to local electricity prices, consumption patterns, and other regional factors. At the international level, macro-level policies such as carbon trading and carbon taxation should be designed in alignment with regional regulations and incentive schemes. Moreover, the RL algorithm for dynamic pricing and energy management proposed in this study is expected to be applied to other domains in the real-time electricity market. The proposed model can be applied to EV charging station operations, which share structural similarities with SSE operations, by adjusting the demand model to reflect consumer behavior. It can also be applied to energy management in smart grids or smart buildings by reformulating the problem as a single-action decision framework focused on cost minimization.

We present several limitations of this study and propose directions for future research. First, this study focuses only on the electricity demand from container vessels using SSE. However, such an energy management strategy could lead to unintended effects, such as energy shortages for other electricity consumers within the port. This study can be extended to a comprehensive energy management system by incorporating other sources of electricity demand, such as reefer containers and port equipment like cranes. Second, while this study addresses the general situation of SSE operations, there are additional factors to consider for real-world application. In some countries, SSE prices are fixed in advance, and the electricity market is not liberalized, making it challenging to earn a profit through operating SSE. In such cases, the algorithm proposed in this study can be further extended to focus on reducing energy costs. Finally, this study focuses on the operational aspect, so the initial cost of the ESS was not considered. Future research from a strategic perspective is needed to analyze economic feasibility by taking into account initial investment costs, operational costs, and revenues. Expanding this research to incorporate such factors would enhance the adoption of SSE operations in real-world settings.

CRediT authorship contribution statement

Chungkwon Oh: Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Ilkyeong Moon:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The authors are grateful for the valuable comments from the area editor and anonymous reviewers. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS–2022–NR066951 and No. RS-2023–00218913).

Appendix A. Myopic algorithm

We introduce the myopic algorithm as a baseline policy for reward shaping. The myopic algorithm is a highly shortsighted approach that makes decisions only to maximize the expected reward of the current step. In the actual research problem, there is uncertainty in renewable energy generation and vessel demand; however, in this algorithm, we implement myopic decisions under a deterministic case. To formalize the myopic decision-making process, we define the following optimization problem, which takes the form of a nonlinear programming (NLP) problem. Consequently, the myopic algorithm solves this problem at each time step t.

$$\max \quad p_t d_t - c_t e_t \tag{A.1}$$
$$\text{s.t.} \quad d_t \leq D_t Prob(f_{n,t} \geq p_t) \tag{A.2}$$
$$d_t \leq \gamma^{disch} I_t + \gamma^{ch}\gamma^{disch}(e_t + r_t) \tag{A.3}$$
$$d_t \leq d^{max} \tag{A.4}$$
$$e_t \leq e^{max} \tag{A.5}$$
$$l \leq p_t \leq u \tag{A.6}$$
$$e_t \geq 0 \tag{A.7}$$
$$d_t \geq 0 \tag{A.8}$$

The objective function (A.1) is identical to Eq. (1). Constraints (A.2) through (A.4) and Constraints (A.5) through (A.8) are derived from Constraints (7)–(9) and Constraints (12)–(13), respectively. In particular, because the port cannot observe $f_{n,t}$ and thus does not know $\lambda_{n,t}$, Constraint (7) is reformulated as Constraint (A.2) based on the available information: the total electricity demand D_t and the distribution of $f_{n,t}$. Under the assumption that the fuel cost follows a uniform distribution with a lower bound l and an upper bound u, it is represented as $d_t \leq D_t(u - p_t)/(u - l)$ using the cumulative distribution function of the uniform distribution. Constraint (A.3) corresponds to Constraint (8) without the term ϵ, because the estimation error ϵ in renewable energy generation is also unobservable.

Appendix B. Time-segmented heuristic algorithm

To provide a benchmark for the proposed RL algorithms, we designed a simple rule-based heuristic algorithm named time-segmented heuristic. It is motivated by the daily pattern in real-time electricity prices. We divide a day into four time segments and define a purchasing rule for electricity in each segment. This algorithm is based on the intuitive principle of purchasing electricity during off-peak periods with lower prices and selling it during peak periods when the prices are higher. The SSE price is determined directly based on the result of the myopic algorithm. The four time segments and corresponding purchasing strategies are as follows:

- Off-peak period (0:00–6:00): Wholesale electricity prices are low. The port purchases electricity in advance for sale during the peak period, and the surplus is stored in the ESS.
- Pre-peak period (6:00–16:00): Wholesale electricity prices gradually increase. Because the energy stored in the ESS is reserved for use during the upcoming peak period, the port adopts a myopic decision-making approach without using the stored energy.
- Peak period (16:00–22:00): Wholesale electricity prices remain relatively high on average. A myopic decision is made based on the amount of energy stored in the ESS during the off-peak period, with the aim of using the stored energy efficiently.
- Post-peak period (22:00–24:00): Wholesale electricity prices gradually decrease. This period aims to consume any remaining stored energy that was not used during the peak period, using a myopic policy that considers the ESS storage level.

Algorithm 3 Time-segmented heuristic algorithm.

```
Input: wholesale price $c_t$, ESS storage level $I_t$, estimated renewable energy generation $r_t$, total electricity consumption $D_t$, current time $T$
Output: purchasing amount $e_t$, SSE price $p_t$
if $T \in$ off-peak period then
    Derive $(e_t, p_t)$ using myopic algorithm with $I_t \leftarrow 0$
    if $c_t \leq \eta^{ch}\eta^{disch} l$ then
        $e_t \leftarrow \min[e^{max},\ 2e_t,\ I^{max} - (I_t + r_t - D_t Prob(f_{n,t} \geq p_t))]$
    end
end
if $T \in$ pre-peak period then
    Derive $(e_t, p_t)$ by myopic algorithm with $I_t \leftarrow 0$
end
if $T \in$ peak period then
    Let $\tau_t^{peak}$ be the number of remaining time steps in the peak period
    Derive $(e_t, p_t)$ by myopic algorithm with $I_t \leftarrow I_t/\tau_t^{peak}$
end
if $T \in$ post-peak period then
    Derive $(e_t, p_t)$ using myopic algorithm
end
```

Algorithm 3 shows the detailed procedure of the heuristic algorithm.

Data availability

The authors do not have permission to share data.

References

[1] UNCTAD, Review of Maritime Transport 2022, Tech. rep., United Nations, 2022.
[2] IMO, Fourth IMO Greenhouse Gas Study, Tech. rep., International Maritime Organization, 2020.
[3] EPA, Shore Power Technology Assessment at u.S. Ports, Tech. rep., United States Environmental Protection Agency, 2022.
[4] L. Zhen, W. Wang, S. Lin, Analytical comparison on two incentive policies for shore power equipped ships in berthing activities, Transp. Res. Part E Logist. Transp. Rev. 161 (2022) 102686.
[5] Y. Gong, Y. Zhou, X. Liu, Y. Huang, Q. Lu, Identifying effective incentive policies for promoting widespread adoption of shore power technology, Transp. Res. Part D Transp. Environ. 126 (2024) 103998.
[6] Z. Zhong, H. Jin, Y. Sun, Y. Zhou, Two incentive policies for green shore power system considering multiple objectives, Comput. Ind. Eng. 194 (2024) 110338.
[7] J. Qi, S. Wang, C. Peng, Shore power management for maritime transportation: status and perspectives, Marit. Transp. Res. 1 (2020) 100004.
[8] K. Kim, Characteristics of economic and environmental benefits of shore power use by container-ship size, J. Mar. Sci. Eng. 10 (5) (2022) 622.
[9] N.B.B. Ahamad, J.M. Guerrero, C.-L. Su, J.C. Vasquez, X. Zhaoxia, Microgrids technologies in future seaports, in: 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), IEEE, 2018, pp. 1–6.
[10] N. Hatziargyriou, Microgrids: Architectures and Control, John Wiley & Sons, 2014.
[11] W. Wang, Y. Peng, X. Li, Q. Qi, P. Feng, Y. Zhang, A two-stage framework for the optimal design of a hybrid renewable energy system for port application, Ocean Eng. 191 (2019) 106555.
[12] A. Molavi, J. Shi, Y. Wu, G.J. Lim, Enabling smart ports through the integration of microgrids: a two-stage stochastic programming approach, Appl. Energy 258 (2020) 114022.
[13] J. Yu, S. Voß, X. Song, Multi-objective optimization of daily use of shore side electricity integrated with quayside operation, J. Clean. Prod. 351 (2022) 131406.
[14] J. Yu, G. Tang, S. Voß, X. Song, Berth allocation and quay crane assignment considering the adoption of different green technologies, Transp. Res. Part E Logist. Transp. Rev. 176 (2023) 103185.
[15] Y. Zhang, C. Liang, J. Shi, G. Lim, Y. Wu, Optimal port microgrid scheduling incorporating onshore power supply and berth allocation under uncertainty, Appl. Energy 313 (2022) 118856.
[16] A. Mao, T. Yu, Z. Ding, S. Fang, J. Guo, Q. Sheng, Optimal scheduling for seaport integrated energy system considering flexible berth allocation, Appl. Energy 308 (2022) 118386.
[17] Ç. Iris, J.S.L. Lam, Optimal energy management and operations planning in seaports with smart grid while harnessing renewable energy under uncertainty, Omega 103 (2021) 102445.
[18] J. Qiu, Y. Tao, S. Lai, J. Zhao, Pricing strategy of cold ironing services for all-electric ships based on carbon integrated electricity price, IEEE Trans. Sustain. Energy 13 (3) (2022) 1553–1565.
[19] K. Yiğit, G. Kökkülünk, A. Parlak, A. Karakaş, Energy cost assessment of shoreside power supply considering the smart grid concept: a case study for a bulk carrier ship, Marit. Policy Manag. 43 (4) (2016) 469–482.
[20] L. Dai, H. Hu, Z. Wang, Y. Shi, W. Ding, An environmental and techno-economic analysis of shore side electricity, Transp. Res. Part D Transp. Environ. 75 (2019) 223–235.
[21] K. Gore, P. Rigot-Müller, J. Coughlan, Cost-benefit assessment of shore side electricity: an Irish perspective, J. Environ. Manag. 326 (2023) 116755.
[22] Y. Wang, W. Ding, L. Dai, H. Hu, D. Jing, How would government subsidize the port on shore side electricity usage improvement?, J. Clean. Prod. 278 (2021) 123893.
[23] Y. Wang, S. Guo, L. Dai, Z. Zhang, H. Hu, Shore side electricity subsidy policy efficiency optimization: from the game theory perspective, Ocean. Coast. Manag. 228 (2022) 106324.
[24] Y. Xing, L. Zhao, R. Huang, Y. Qian, Green energy subsidy structure design under the impact of conventional energy price uncertainty, Comput. Ind. Eng. 174 (2022) 108798.
[25] Y.-T. Peng, Y. Wang, Z.-C. Li, D. Sheng, Subsidy policy selection for shore power promotion: Subsidizing facility investment or price of shore power?, Transp. Policy 140 (2023) 128–147.
[26] S. Bai, N. Xu, The online seller's optimal price and inventory policies under different payment schemes, Eur. J. Ind. Eng. 10 (3) (2016) 285–300.

[27] H. Qin, D. Simchi-Levi, L. Wang, Data-driven approximation schemes for joint pricing and inventory control models, Manag. Sci. 68 (9) (2022) 6591–6609.

[28] Q. Zhou, Y. Yang, S. Fu, Deep reinforcement learning approach for solving joint pricing and inventory problem with reference price effects, Expert Syst. Appl. 195 (2022) 116564.

[29] T. Yavuz, O. Kaya, Deep reinforcement learning algorithms for dynamic pricing and inventory management of perishable products, Appl. Soft Comput. 163 (2024) 111864.

[30] J. Kumar, A.A. Memon, L. Kumpulainen, K. Kauhaniemi, O. Palizban, Design and analysis of new harbour grid models to facilitate multiple scenarios of battery charging and onshore supply for modern vessels, Energies 12 (12) (2019) 2354.

[31] Q. Zhang, J. Qi, L. Zhen, Optimization of integrated energy system considering multi-energy collaboration in carbon-free hydrogen port, Transp. Res. Part E Logist. Transp. Rev. 180 (2023) 103351.

[32] P. Ge, D. Tang, Y. Yuan, J.M. Guerrero, E. Zio, A hierarchical multi-objective co-optimization framework for sizing and energy management of coupled hydrogen-electricity energy storage systems at ports, Appl. Energy 384 (2025) 125451.

[33] D. Srinivasan, S. Rajgarhia, B.M. Radhakrishnan, A. Sharma, H.P. Khincha, Game-theory based dynamic pricing strategies for demand side management in smart grids, Energy 126 (2017) 132–143.

[34] R. Lu, S.H. Hong, X. Zhang, A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach, Appl. Energy 220 (2018) 220–230.

[35] Y. Zhang, Q. Yang, D. Li, D. An, A reinforcement and imitation learning method for pricing strategy of electricity retailer with customers, flexibility, Appl. Energy 323 (2022) 119543.

[36] M.J. Alves, C.H. Antunes, A semivectorial bilevel programming approach to optimize electricity dynamic time-of-use retail pricing, Comput. Oper. Res. 92 (2018) 130–144.

[37] X. Dong, Y. Mu, X. Xu, H. Jia, J. Wu, X. Yu, Y. Qi, A charging pricing strategy of electric vehicle fast charging stations for the voltage control of electricity distribution networks, Appl. Energy 225 (2018) 857–868.

[38] D. Qiu, Y. Ye, D. Papadaskalopoulos, G. Strbac, A deep reinforcement learning method for pricing electric vehicles with discrete charging levels, IEEE Trans. Ind. Appl. 56 (5) (2020) 5901–5912.

[39] Z. Zhao, C.K.M. Lee, Dynamic pricing for EV charging stations: a deep reinforcement learning approach, IEEE Trans. Transp. Electric. 8 (2) (2021) 2456–2468.

[40] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, H. Hashimoto, Deep reinforcement learning for joint bidding and pricing of load serving entity, IEEE Trans. Smart Grid 10 (6) (2019) 6366–6375.

[41] H. Xu, Q. Wu, J. Wen, Z. Yang, Joint bidding and pricing for electricity retailers based on multi-task deep reinforcement learning, Int. J. Electr. Power Energy Syst. 138 (2022) 107897.

[42] C. Luo, Y.-F. Huang, V. Gupta, Stochastic dynamic pricing for EV charging stations with renewable integration and energy storage, IEEE Trans. Smart Grid 9 (2) (2017) 1494–1505.

[43] S. Lee, D.-H. Choi, Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: a privacy-preserving deep reinforcement learning approach, Appl. Energy 304 (2021) 117754.

[44] P. Rajan, S. Jeevananthan, An adjustable gain three port converter for battery and grid integration in remote location microgrid systems, Renew. Energy 179 (2021) 1404–1423.

[45] S. Wen, T. Zhao, Y. Tang, Y. Xu, M. Zhu, S. Fang, Z. Ding, Coordinated optimal energy management and voyage scheduling for all-electric ships based on predicted shore-side electricity price, IEEE Trans. Ind. Appl. 57 (1) (2020) 139–148.

[46] S. Ghosh, L.H. Lee, S.H. Ng, Bunkering decisions for a shipping liner in an uncertain environment with service contract, Eur. J. Oper. Res. 244 (3) (2015) 792–802.

[47] T.P. Lillicrap, Continuous control with deep reinforcement learning, arXiv preprint arXiv:1509.02971, 2015.

[48] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, M. Riedmiller, Deterministic policy gradient algorithms, in: International Conference on Machine Learning, PMLR, 2014, pp. 387–395.

[49] S. Fujimoto, H. Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: International Conference on Machine Learning, PMLR, 2018, pp. 1587–1596.

[50] H. Hasselt, Double q-learning, Adv. Neural Inf. Process. Syst. 23 (2010).

[51] A.Y. Ng, D. Harada, S. Russell, Policy invariance under reward transformations: theory and application to reward shaping, in: ICML, vol. 99, 1999, pp. 278–287.

[52] B.J. De Moor, J. Gijsbrechts, R.N. Boute, Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management, Eur. J. Oper. Res. 301 (2) (2022) 535–545.

[53] J. Lee, Y. Shin, I. Moon, A hybrid deep reinforcement learning approach for a proactive transshipment of fresh food in the online–offline channel system, Transp. Res. Part E Logist. Transp. Rev. 187 (2024) 103576.

[54] R. Lu, Z. Jiang, H. Wu, Y. Ding, D. Wang, H.-T. Zhang, Reward shaping-based actor–critic deep reinforcement learning for residential energy management, IEEE Trans. Ind. Inf. 19 (3) (2022) 2662–2673.

[55] V. Mnih, Asynchronous methods for deep reinforcement learning, arXiv preprint arXiv:1602.01783, 2016.

[56] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables, vol. 55, Courier Corporation, 1965.

[57] J. Ding, W. Chen, Effects of trade credit insurance on remanufacturing decisions under carbon tax and emissions abatement, Eur. J. Ind. Eng. 16 (6) (2022) 679–705.

[58] E.K. Tajani, A.G. Kanafi, M. Daneshmand-Mehr, A.-A. Hosseinzadeh, Designing the agile green sustainable multi-channel closed-loop supply chain with dependent demand to price and greenness under epistemic uncertainty, Eur. J. Ind. Eng. 18 (4) (2024) 557–605.