Received January 10, 2020, accepted January 20, 2020, date of publication January 23, 2020, date of current version February 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968841

Reinforcement-Learning-Based Energy Storage System Operation Strategies to Manage Wind Power Forecast Uncertainty

EUNSUNG OH[1], (Member, IEEE), AND HANHO WANG[2], (Member, IEEE)

[1]Department of Electrical and Electronic Engineering, Hanseo University, Chungcheongnam 31962, South Korea
[2]Department Smart Information and Telecommunication Engineering, Sangmyung University, Chungcheongnam 31066, South Korea

Corresponding author: Hanho Wang (hhwang@smu.ac.kr)

This was supported by the National Research Foundation of Korea grant funded by the Korean Government (Ministry of Science and ICT) under Grant 2017R1E1A1A03070136.

ABSTRACT Currently, renewable-energy-based power generation is rapidly developing to tackle climate change; however, the use of renewable energy is limited owing to the uncertainty related to renewable energy sources. In particular, energy storage systems (ESSs), which are critical for implementing wind power generation (WPG), entail a wide uncertainty range. Herein, a reinforcement leaning (RL)-based ESS operation strategy is investigated for managing the WPG forecast uncertainty. First, a WPG forecast uncertainty minimization problem is formulated with respect to the ESS operation, subject to ESS constraints, and then, the problem is presented as a Markov decision process (MDP) model, with the state-action space limited by the ESS characteristics. To achieve the optimal solution of the MDP model, an expected state–action–reward–state–action (SARSA) method, which is robust toward the dispersion of the system environment, is employed. Further, frequency-domain data screening based on the k-mean clustering method is implemented to improve learning performance by reducing the variance of the WPG forecast uncertainty. Extensive simulations are conducted based on practical WPG generation data and forecasting. Results indicate that the proposed clustered RL-based ESS operation strategy can manage the WPG forecast uncertainty more effectively than conventional Q-learning-based methods; additionally, the proposed method demonstrates a near-optimal performance within a 1%-point analysis gap to the optimal solution, which requires complete information, including future values.

INDEX TERMS Energy storage, forecasting, Markov decision process, mean absolute error, reinforcement learning, reliability, renewable, uncertainty, wind power.

NOMENCLATURE

$\mathcal{T}$	Operation time horizon, i.e., $\mathcal{T} = \{1, \cdots, t, \cdots, T\}$
ΔT	Operation time interval [h]
g_t	Actual wind power generation at time t
$\hat{g}_t$	Wind power generation forecasting at time t
e_t	Forecast error at time t, i.e., $e_t = \hat{g}_t - g_t$

ENERGY STORAGE SYSTEM

a_t	Action at time t
q_t	Charge/discharge quantity at time t
c_t	State of charge at time t
C_{PS}	Power subsystem (PS) capacity [kW]
C_{ES}	Energy subsystem (ES) capacity [kWh]
η_{PS}	Compensation factor for PS within (0, 1]
η_{ES}	Compensation factor for ES within (0, 1)

OPERATION STRATEGY

s_i^t	State i at stage t
S_t	Available state set at stage t, i.e., $s_i^t \in S_t$
a_j^t	Action j at stage t
A_t	Feasible action set at stage t, i.e., $a_j^t \in A_t$
r_t	Reward at stage t
R_t	Return at stage t, i.e., $R_t = r_t + \gamma R_{t+1}$
γ	Discount factor within (0, 1]
α	Learning rate within (0, 1]

The associate editor coordinating the review of this manuscript and approving it for publication was Eklas Hossain.

$Q(s_t, a_t)$	State-action value function
K	Number of data clusters
$\boldsymbol{c}$	Set of data clusters, i.e., $c = \{c_1, \cdots, c_k, \cdots, c_K\}$

I. INTRODUCTION

In 2018, global primary energy consumption increased by 2.9%, which is the highest rate of primary energy consumption over the last 10 years; in addition, the global power demand increased by 3.7% [1]. On the supply side, the increase in renewable-energy-based power generation was 14.5%, which significantly contributed to the overall increase in energy consumption. Owing to environmental problems such as climate change, a continuous increase in renewable-energy-based power generation is expected [2]. In particular, the globally produced wind energy corresponds to approximately 20% of the electricity power generated using renewable energies [3].

An increase in renewable-energy-based power generation decreases the power grid stability [4], [5]. Although various power forecasting methods such as curve modeling [6], the multimodel combination approach [7], vector autoregressive model [8], and neural networks [9] were researched, uncertainty cannot be completely eliminated owing to the intermittent and fluctuating nature of renewable energy resources.

Energy storage systems (ESSs) are critical for the management of wind power generation (WPG) forecast uncertainties [10]. The fundamental role of an ESS is the charging of chemical, physical, or electrical materials with surplus energy and the discharging of energy according to the operational objective. Battery energy storage systems (BESSs) were recently considered because of their convenient control and operational efficiency [11]. To better schedule the ramping capacity of the ESS and a generation unit, a continuous time method based on coefficients of the Bernstein polynomial is proposed herein that provides, compared to existing approaches, a more accurate representation of the sub-hourly ramping needs following fast sub-hourly ramping of WPG [12]. A systematic approach to evaluate the level of flexibility of a power system by unequivocally considering fast-ramping units, hourly demand response and energy storage is provided that is considered a flexibility index to evaluate the system's technical aptitude [13]. A stochastic optimization framework to coordinate the flexibility resources dealing with the uncertainty of WPGs and equipment failures is formulated as mixed-integer linear programming [14]. ESSs are implemented in various applications for wind power generation, such as frequency regulation, peak shaving, and ancillary services [15]–[17]. To use ESSs for the management of WPG forecast uncertainty, various approaches have been researched in the literature—meta heuristic-based approaches such as the genetic algorithm (GA) [18], particle swarm optimization (PSO) [19], mixed hybrid algorithm approaches [20], scenario-based stochastic approaches [21], and discrete Fourier transform (DFT)-based approaches [22]. However, owing to the recursion of the ESS operation, a generalized methodology is required.

This study focuses on a reinforcement learning (RL)-based ESS operation. RL is useful for generalization as it enables the design of model-free approaches [23]. Hence, in recent years, various studies have been conducted using RL for energy management. For energy-efficient electric vehicle management, an RL-based velocity predictive energy management strategy [24] and an RL-based real-time energy management approach were employed to minimize the energy loss of the ESS in a plug-in hybrid electric vehicle [25]. For an adaptive demand response (DR), a fully automated energy management scheduling was formulated as an RL problem, and then solved by decomposing the problem over device clusters [26]. Moreover, RL is used as the decision-making framework for dynamic pricing-based demand response programs [27]. Several RL-based approaches for DR are summarized in [28]. In addition, RL-based energy management algorithms were implemented in microgrid (MG) environments to maximize the self-consumption of local photovoltaic production [29], supplier and consumer profits [30], [31], the utilization of a community ESS [32], [33], and energy trading among MG to increase utilization [34]. Furthermore, RL-based management algorithms can be implemented in various applications such as smart building energy management [35], power smoothing [36], prosumer energy trading [37], [38], and electricity market trading [39] in addition to operations and the maintenance of power grids [40]. Majority of the previous research was based on Q-learning. This is because Q-learning-based algorithms that directly approximate a value function as the optimal action-value function simplify algorithm implementation and enable early convergence [41]. However, typically, these algorithms cannot converge to optimal strategies because of the perturbation of parameters [42].

Herein, an RL-based ESS operation strategy is proposed for the management of WPG forecast uncertainty. The ESS operation was modeled using a Markov Decision Process (MDP), while considering its operational characteristics. Using the model, an expected state–action–reward–state–action (SARSA)-based ESS operation strategy was employed for managing the WPG forecast uncertainty. An expected SARSA is more robust to the dispersion of the system environment rather than the Q-learning method; hence, it is suitable for the management of WPG forecast uncertainty. Moreover, to reduce the perturbation of parameters, the proposed ESS operation strategy combines the expected SARSA with frequency domain data clustering. The WPG forecast uncertainty is related to the frequency domain characteristics of WPG forecasting [22]. Screening data using frequency domain forecasting data clustering improves the effectiveness of the proposed ESS operation strategy by reducing learning variability.

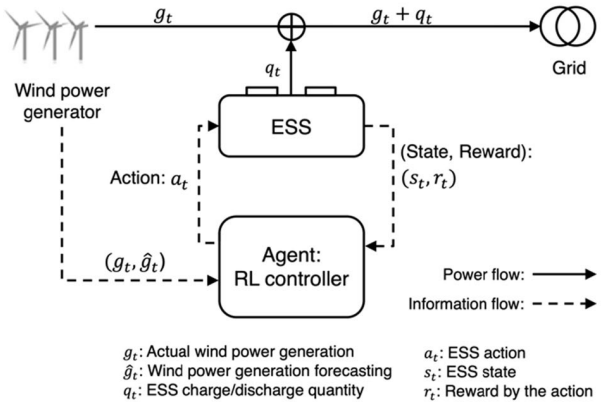


FIGURE 1. Wind power generation (WPG) system model with an energy storage system (ESS).

The remainder of this article is organized as follows. Section II describes system models and problem formulation of the ESS operation. Section III discusses the design of the proposed strategy. Section IV presents measurement studies using practical WPG generation, and its forecasting data is applied to the proposed strategy. Finally, Section V presets the study's conclusion.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. UNCERTAINTY MODEL

In this study, a grid-connected WPG system was considered. Under the assumption that g_t and $\hat{g}_t$ represent the actual WPG and its forecasting at time t, respectively, the WPG forecast uncertainty at time t is defined as

$$e_t = \hat{g}_t - g_t. \tag{1}$$

An ESS was added to reduce the uncertainty, as shown in Figure 1. With the inclusion of the ESS operation, which involves charging or discharging energy q_t, the uncertainty can be calculated as follows:

$$\epsilon_t = e_t - q_t. \tag{2}$$

The objective of the ESS operation is to eliminate the uncertainty in (2). Therefore, the positive action of q_t expresses the discharge operation to grid and vice versa.

B. ENERGY STORAGE SYSTEM MODEL

The ESS contains a power subsystem (PS) and an energy subsystem (ES) [22]. The PS constructed as the power conversion system (PCS) limits the maximum instantaneous charging and discharging power, and the energy stored in the ES determines the ESS service time. Therefore, the ESS operation should be performed within these two constraint regions.

First, the ESS charging or discharging action at each decision time a_t is constricted by the maximum PS capacity C_{PS},

$$-C_{PS} \le a_t \le C_{PS}, \quad \forall t \in \mathcal{T}, \tag{3}$$

where $\mathcal{T}$ is the ESS operation time horizon, i.e., $\mathcal{T} = \{1, \cdots, t, \cdots, T\}$. Considering that the PS efficiency $\eta_{PS} \in (0, 1]$, the actual operation quantity of PS q_t can be measured as follows:

$$q_t = \begin{cases} \eta_{PS} a_t, & \text{if } a_t \ge 0, \\ a_t / \eta_{PS}, & \text{if } a_t < 0. \end{cases} \tag{4}$$

Second, ESS action is operated in the stored energy range, which is referred to as the state-of-charge (SoC). The SoC at time t, represented by c_t, can be expressed as follows:

$$c_t = c_{t-1} + q_t \Delta T, \tag{5}$$

where ΔT is the operation time interval.

The ES capacity C_{ES} limits the SoC, i.e., the accumulated ESS action, as follows:

$$C_{ES}^{\min} \le c_t \le C_{ES}^{\max}. \tag{6}$$

Here, $C_{ES}^{\min}$ and $C_{ES}^{\max}$ represent the minimum and maximum operable ES capacity ranges, respectively, under the consideration of the depth of discharge. Similar to the PS compensation factor, $C_{ES}^{\min} = 0$ and $C_{ES}^{\max} = \eta_{ES} C_{ES}$ can be obtained using the ES compensation factor (efficiency) $\eta_{ES} \in (0, 1]$.

C. PROBLEM FORMULATION

The aim of this study was to determine ESS operation action for WPG forecast uncertainty management. The mean absolute error (MAE) was used as the uncertainty management performance metric.

During the ESS operation time horizon, the MAE was calculated as follows:

$$\mathcal{O}(\boldsymbol{a}) = \frac{1}{T} \sum\nolimits_{t \in \mathcal{T}} |e_t - q_t| = \frac{1}{T} \sum\nolimits_{t \in \mathcal{T}} |\epsilon_t|, \tag{7}$$

where $\boldsymbol{a} = \{a_1, \cdots, a_t, \cdots, a_T\}$.

With the ESS operation constraints, the WPG forecast uncertainty management problem solved by the ESS operation can be expressed as follows:

$$\begin{aligned} &\min_{\boldsymbol{a}} \quad \mathcal{O}(\boldsymbol{a}) \\ &\text{subject to} \ \ (3) \text{ and} (6). \end{aligned} \tag{8}$$

With complete information, including the WPG on forward time, the problem in (8) can be solved using iteration-based search algorithms such as the gradient descent method and the Newton method [43]. However, this assumption is not in accordance with causality; thus, it cannot be implemented in the real world [44]. However, for the performance comparison with the proposed ESS operation strategy, the solution of this problem based on complete information was considered as the optimal solution.

III. RL-BASED ESS OPERATION STRATEGY

A. MARKOV DECISION PROCESS

The ESS operation for the management of the WPG forecast uncertainty is a sequential decision-making (SDM) problem, as expressed in (8). The Markov decision process (MDP) is a classical formalization of SDM, and it is an idealized mathematical form of the RL problem [23]. To meet the

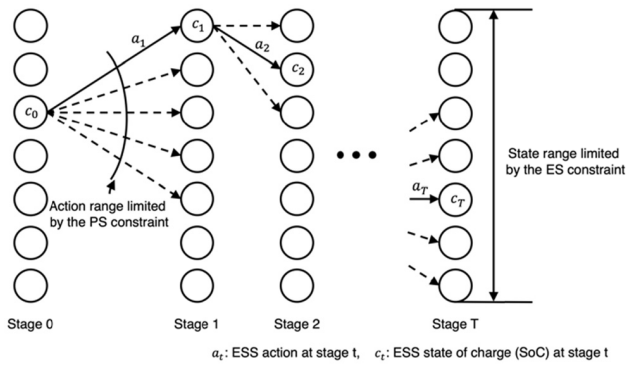


FIGURE 2. A state–action space model for ESS operation. State and action ranges are limited by the ESS condition, i.e., ES and PS constraints.

optimal criteria under an MDP model, a state–action space and transaction probability among states are required. However, an RL approach reduces the constraint of the requirement of the transaction probability among states. Therefore, a state–action space is discussed in this section.

The ESS is operated during the operation time horizon $\mathcal{T}$ with T decision epochs. Therefore, the state–action space for the ESS operation has $T + 1$ decision stages, which include the initial stage, as shown in Figure 2. Moreover, stages and actions express the ESS conditions, i.e., the SoC and the ESS operation, respectively.

At each stage, the state indicates that the SoC at the stage is limited by the ES capacity, and all available states at stage t can be expressed as follows:

$$S_t = \{s_1^t, s_2^t, \cdots, s_i^t, \cdots, s_{\kappa_s}^t\}, \tag{9}$$

where $\kappa_s = \lfloor (C_{ES}^{\max} - C_{ES}^{\min})/\delta \rfloor$ with respect to the ES capacity constraint expressed by (6), and $\lfloor \cdot \rfloor$ and δ represent the floor operation and unit action step of the ESS operation, respectively. The MDP-based RL approach is only solved under the discrete condition; thus, discrete ESS operation is required. Discrete ESS operation generates quantization error; however, the error is bound according to the step size [45].

Similar to the state, all operable actions are restricted by the PS capacity, as follows:

$$A_0 = \left\{a_{-\kappa_a}^0, \cdots, a_j^0, \cdots, a_{\kappa_a}^0\right\}, \tag{10}$$

where $\kappa_a = \lfloor C_{PS}/\delta \rfloor$, in accordance with the PS capacity constraint expressed by (3). Moreover, the action at each stage should be determined within the state range expressed by (9). The next state is then determined by the current state and the current action set, as follows:

$$s_l^{t+1} \leftarrow < s_i^t, a_j^t > . \tag{11}$$

Hence, the feasible action range at stage t can be expressed as follows:

$$A_t = \{a_{j_{\min}}^t, \cdots, a_j^t, \cdots, a_{j_{\max}}^t, \tag{12}$$

where $j_{\min} = \max(-\kappa_a, 1 - i)$ and $j_{\max} = \min(\kappa_a, \kappa_s - i)$.

Figure 2 presents an example of the state–action model for ESS operation when $\kappa_s = 7$ and $\kappa_a = 2$. The state at Stage 1 is $s_7^1 (= s_{\kappa_s}^1)$; thus, the feasible action range is $A_1 = \{a_{-2}^1, a_{-1}^1, a_0^1\}$. With $a_1 = a_{-1}^1$ selected as the action at Stage 1, the state at Stage 2 is s_6^2, and it can be expressed as follows:

$$s_2 = s_6^2 \leftarrow < s_1, \quad a_1 > = < s_7^1, a_{-1}^1 > . \tag{13}$$

B. RL-OPTIMAL POLICY

An RL-based ESS operation involves the decision-making of the action at each stage, under the consideration of the current state and the feasible action range, as presented by the previous MDP model.

The goal of the ESS operation is the minimization of the WPG forecast uncertainty, which is represented by the MAE during the ESS operation time horizon, as shown in (7). Therefore, the objective function at decision stage t can be expressed as follows:

$$\begin{aligned} \mathcal{O}_t(a_t|e_t) &= \frac{1}{T}\sum\nolimits_{i=t}^{T} |\epsilon_t| \\ &= \frac{1}{T}|\epsilon_t| + \frac{1}{T}\sum\nolimits_{i=t+1}^{T} |\hat{\epsilon}_i| \\ &= \frac{1}{T}|\epsilon_t| + \mathcal{O}_{t+1}(\hat{a}_{t+1}|\hat{e}_{t+1}), \end{aligned} \tag{14}$$

where the values with hats represent the expected values.

In RL, the objective function is modeled as the reward and return. The reward represents the instantaneous value from the action at each decision stage according to the environment and current state, as shown in Figure 1; additionally, the return represents the cumulative reward time t onward. With the reward at decision stage t represented by r_t, it can be expressed as the forecast uncertainty with the ESS operation at decision stage t:

$$r_t = \frac{1}{T}|\epsilon_t|. \tag{15}$$

Moreover, the return R_t is defined using the reward r_t, as follows:

$$\begin{aligned} R_t &= r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t} r_T \\ &= r_t + \gamma R_{t+1}, \end{aligned} \tag{16}$$

where γ is the discount factor in $(0, 1]$, which reduces the risk of the expected value from the onward decision time. Subsequently, the return in (16) is the discounted objective function in (14).

To design the decision-making of the action, the state–action value function is defined, which expresses the performance of a determined action at a given state, as follows:

$$\begin{aligned} Q(s_t, a_t) &= \mathbb{E}[R_t|s_t, a_t] \\ &= \mathbb{E}[r_t + \gamma Q(s_{t+1}, a_{t+1})|s_t, a_t]. \end{aligned} \tag{17}$$

A policy π is a decision-making strategy of the action. It is expressed as the transaction probability of an action, given the state at each decision stage, i.e., $\pi = \Pr(a_t|s_t), \forall t \in \mathcal{T}$,

$s_t \in S_t, a_t \in A_t$. The optimal policy is the strategy implemented for the minimization of the state–action value of all states, $\pi^* = \text{argmin}_\pi Q(s_t, a_t), \forall t \in \mathcal{T}, s_t \in S_t, a_t \in A_t$. With respect to the state–action value function, it can be expressed as $Q^*(s_t, a_t) = \min_\pi Q_\pi(s_t, a_t), \forall t \in \mathcal{T}, s_t \in S_t, a_t \in A_t$, where $Q_\pi(s_t, a_t)$ expresses the state-action value when the policy π is applied. Based on (17), the Bellman optimality equation for the state–action function can be expressed as follows [46]:

$$\begin{aligned} Q^*(s_t, a_t) &= \mathbb{E}[r_t + \gamma \min_{a_{t+1} \in A_{t+1}} Q(s_{t+1}, a_{t+1}) | s_t, a_t] \\ &= \mathbb{E}[r_t + \gamma Q^*(s_{t+1}, a_{t+1}) | s_t, a_t]. \end{aligned} \tag{18}$$

The optimal state–action value function in (18) indicates that the optimal policy is based on the decision of the local optimal action at each decision state t, given that the expected reward from the onward decision time is taken care of the optimal value. Therefore, the optimal action is determined as follows:

$$\begin{aligned} a_t^* &= \arg\min_{a_t \in A_t} Q^*(s_t, a_t) \\ &= \arg\min_{a_t \in A_t} \mathbb{E}[r_t | s_t, a_t] + \gamma Q^*(s_{t+1}, a_{t+1}). \end{aligned} \tag{19}$$

If the state–action probability at each decision stage is known, the optimal action in (19) is determined based on the calculation of the optimal state–action function in (18) using dynamic programming [46]. However, an impractical method can be employed. In this study, the optimal policy is learned by estimating the optimal state–value function using model-free RL methods.

Owing to its simplicity, the Q-learning-based RL method is widely used [24]–[40]. In the Q-leaning-based RL method, the optimal action is estimated as follows:

$$a_t^{QL} = \arg\min_{a_t \in A_t} \left[r_t + \gamma \min_{a_{t+1} \in A_{t+1}} Q^{QL}(s_{t+1}, a_{t+1}) \right]. \tag{20}$$

Moreover, the state–action value function is updated as follows:

$$\begin{aligned} Q^{QL}(s_t, a_t) \leftarrow (1-\alpha) Q^{QL}(s_t, a_t) \\ + \alpha \left[r_t + \gamma \min_{a_{t+1} \in A_{t+1}} Q^{QL}(s_{t+1}, a_{t+1}) \right], \end{aligned} \tag{21}$$

where α is a learning rate of convergence in (0, 1].

However, the WPG forecast uncertainty contains significantly high variances, which increase throughout the decision time horizon [22]. This reduces the reliability of the expected value from the onward decision time, i.e., $Q^{QL}(s_{t+1}, a_{t+1})$. Consequently, the wrong decision is made in (20), and the convergence speed of the state-action value function is reduced to the optimal function in (21).

The expected SARSA-based RL method is more robust to the variance of the state-action value from the onward decision time [23]. Considering the mean of the state–action value, the method reduces the sensitivity of the expected values. In the expected SARSA-based RL method, the action is determined as follows:

$$a_t^{ES} = \arg\min_{a_t \in A_t} \left[r_t + \gamma \mathbb{E}_{A_{t+1}} \left\{ Q^{ES}(s_{t+1}, a_{t+1}) \right\} \right]. \tag{22}$$

The state-action value function is updated as follows:

$$\begin{aligned} Q^{ES}(s_t, a_t) \leftarrow (1-\alpha) Q^{ES}(s_t, a_t) \\ + \alpha \left[r_t + \gamma \mathbb{E}_{A_{t+1}} \left\{ Q^{ES}(s_{t+1}, a_{t+1}) \right\} \right]. \end{aligned} \tag{23}$$

C. FREQUENCY DOMAIN DATA CLUSTERING

The expected SARSA is the approach employed for the reduction of the risk due to the variance of the uncertainty during the operation process. However, the data pre-process is an effective technique that can be employed to increase the learning performance [47].

The WPG forecasting accuracy, which determines the WPG forecasting performance, is related to the gradient of the time-series data [22]. This is because the WPG forecasting algorithm cannot easily track the instantaneous changes of the data. Therefore, the frequency-domain analysis is an effective method for the characterization of the WPG forecast uncertainty [48]. The time-series WPG forecasting data is converted to a frequency-domain sequence using DFT, as follows:

$$G_j = \sum_{t=1}^{T} \hat{g}_t e^{-j2\pi(t-1)(j-1)/T}, \quad j = \{1, \cdots, J\}, \tag{24}$$

where j is the frequency element with the same length as the time-series WPG forecasting sequence, $J = T$, and $G = \{G_1, \cdots, G_j, \cdots, G_J\}$.

For the data pre-processing, a k-means clustering technique was employed. The k-mean clustering algorithm is a vector quantization method for the classification of data into K clusters [49]. Mathematically, it is formulated for the determination of sets $c = \{c_1, \cdots, c_k, \cdots, c_K\}$,

$$\arg\min_{c} \sum_{k=1}^{K} \sum_{G \in c_k} ||G - \mu_k||_2, \tag{25}$$

where $|| \cdot ||_2$ expresses the Euclidean norm operation, and μ_k represents the mean points with T-dimensional space in c_k. The problem is a type of NP-hard problem [50]; however, it can be solved using the Lloyd algorithm, which repetitively determines the centroids of Voronoi diagrams [49].

D. CLUSTERED RL-BASED ESS OPERATION STRATEGY

The proposed strategy is an expected SARSA-based ESS operation method that combines frequency-domain WPG forecasting data clustering, i.e., a clustered RL-based ESS operation strategy. Furthermore, it includes data preprocessing and optimal policy learning, as follows:

In the proposed clustered RL-based ESS operation algorithm, the WPG forecasting data is first clustered under the consideration of the frequency-domain characteristic (Steps 1–7). To apply the k-means clustering method, the number of clusters and mean points of the clusters should be determined. Given that the number of Q-tables is determined according to the number of clusters, the cluster number is determined according to the memory condition of the system in Step 2. In addition, the mean points of the clusters can be determined using the historical WPG forecasting data in Step 3. By converting the time-series WPG forecasting data

Algorithm 1 A Clustered RL-Based ESS Operation Algorithm

Datapreprocessing

1: *Initialization*
2: Set a number of clusters K.
3: Train mean points μ_k using historical WPG forecasting data.
4: *Data clustering*
5: Convert to a frequency-domain sequence G using (24).
6: Set cluster k as $k = \text{argmin}_k ||G - \mu_k||_2$.
7: Update μ_k including G.

Optimal policy learning

8: *Initialization*
9: Set Q^{ES} as Q_k from $Q = \{Q_1, \cdots, Q_K\}$.
10: Set $s_1 \leftarrow c_0$ and A_1 using (12).
11: *Policy learning*
12: **For** $t = \{1, \cdots, T\}$,
13: Set a_t^{ES} in A_t using (22).
14: Update s_{t+1}, A_{t+1}, and Q^{ES} using (11) and (23).
15: **end for**

to a frequency-domain sequence in Step 5, the data cluster containing the minimum Euclidean distance with respect to the mean points of the clusters in Step 6 is set. Thereafter, the cluster selected on the basis of the mean point is updated considering the data in Step 7.

In the optimal policy learning process, the k-th Q-table is loaded according to the selected cluster k in Step 9. The initial state s_1 is considered as the current ESS condition c_0, and the ESS action range is determined based on the state and ESS characteristic in Step 10. The policy learning is processed during the ESS operation time horizon $\mathcal{T}$ (Steps 12–15). The optimal action is selected to minimize the expected reward (mean absolute error) in Step 13. The next state, which is the action range of the next state and the Q-table, is updated according to the selected optimal action in Step 14. When the stage is the terminal stage, i.e., $t = T$; the Q-value of the next stage is set to zero in Steps 13 and 14.

The procedure of the proposed clustered RL-based ESS operation algorithm is presented in Figure 3 as a flowchart.

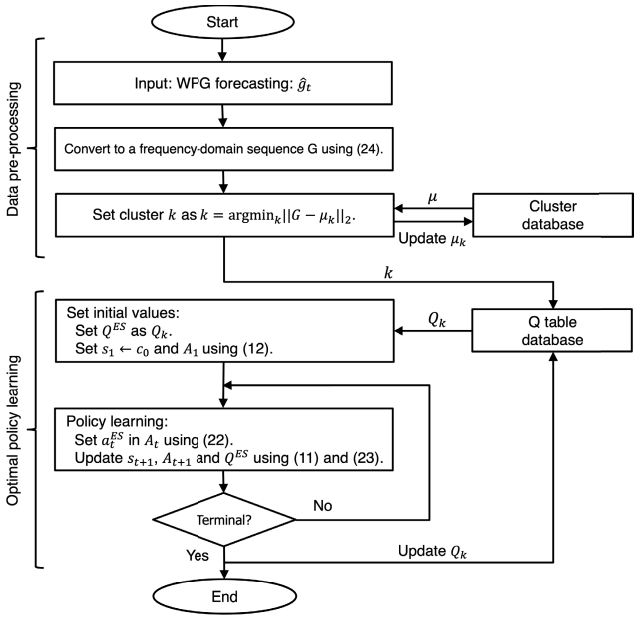


FIGURE 3. Flowchart of the proposed algorithm.

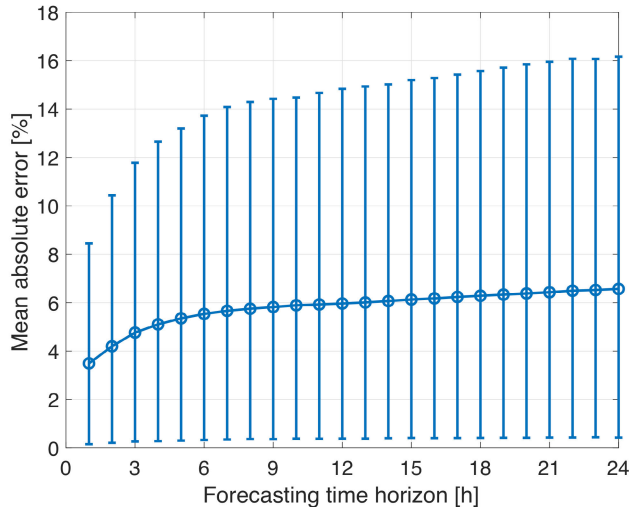


FIGURE 4. MAE of the BPA WPG forecasting with respect to forecasting time horizon.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL ENVIRONMENT

The results of this study were evaluated using WPG and its forecasting data recorded by the Bonneville Power Administration (BPA), United States Department of Energy. The BPA is a non-profit federal power marketing administration based in the Pacific Northwest [51]. The BPA territory includes Idaho, Oregon, Washington, western Montana, and small portions of eastern Montana, California, Nevada, Utah, and Wyoming. The cumulative WPG capacity in the BPA balancing authority area was 4782 MW from 2015–2017, which decreased to 2764 MW in July 2018. Therefore, herein, the data obtained for 360 days in 2017 was used.

The MAE of the BPA WPG forecasting increased according to the forecasting time horizon, as shown in Figure 4. In particular, the variance of the forecasting error increased, as indicated by the bar in the figure that presents the error range from the 10%–90% quantiles of the total error range. The error dispersion increases the difficulty of the ESS operation. Therefore, this study reveals the performance with respect to the changes in the forecasting time horizon. For a generic explanation, the results are presented as the value related to the WPG capacity; thus, quantities are expressed in per-unit (p.u.).

Lithium-ion battery systems are employed as the ESSs in various applications [14]. The characteristics of the ESS were set as $\eta_{PS} = 0.95$ and $\eta_{ES} = 0.9$, under the assumption of a

90% round trip efficiency and 10% depth of discharge (DoD) margin. The ESS size was assumed as 1 per-unit (p.u.), and the service time was 2 h, i.e., the charging rate (C-rate) was 0.5. However, a discussion on the performance with respect to the size is presented here.

The WPG forecast uncertainty exhibited significant variance, and the MAE was calculated with equal priority over the entire operation time duration. Therefore, for the policy leaning process, the discount factor and learning rate were set to $\gamma = 0.95$ and $\alpha = 0.1$, respectively.

Simulations in this study were implemented on a 64-bit PC with a 3 GHz 8-Core Intel Xeon E5 CPU and 64 GB RAM, using MATLAB R2018a with an IBM CPLEX optimization studio.

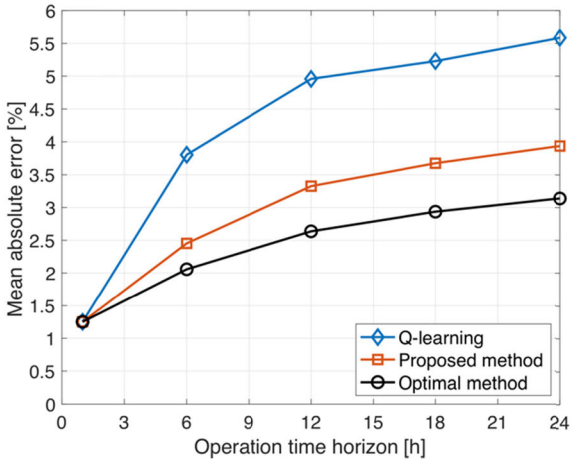
B. PERFORMANCE OF MAE

MAE performance was compared with that of the optimal ESS operation method, which required complete information on future values, as formulated in (8) in addition to the Q-learning based algorithm presented in (20) and (21) that was applied in majority of prior research. In the proposed method, three clusters, i.e., $K = 3$, were considered.

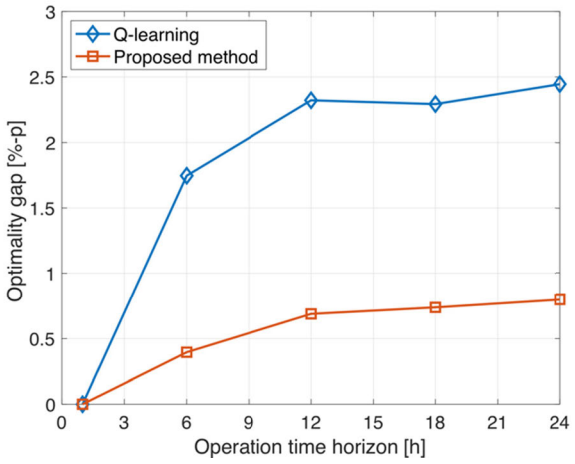
Figure 5(a) presents the MAE with ESS operations. The blue line with the diamond shapes, the red line with the square shapes, and the black line with the circular shapes indicate the results of the application of Q-learning, the proposed method, and the optimal method, respectively. The MAE was found to increase in accordance with an increase in the operation time horizon. This is because the WPG forecasting error and its variance increased when the operation time horizon was long, as shown in Figure 4. The performance of the proposed method was found to be superior to that of the Q-learning-based method, and the same trend as the results of the optimal method was observed. As shown in Figure 5(b), the optimality gap between the MAE based on the Q-learning-based method and the optimal method was approximately 2.4%-point; however, the proposed method exhibited an optimality gap of less than 0.8%-point optimality gap. This is because the proposed method appropriately managed the dispersion of the WPG forecasting error based on the expectation and clustering.

C. EFFECT OF METHODOLOGY

Figure 6 presents a comparison of the MAE according to a combination of methodologies. The blue lines with diamond shapes and red lines with square shapes indicate the results of the application of the Q-learning and expected SARSA-based methods, respectively. The dashed and solid lines indicate the results of the cases without and with clustering, respectively. For the Q-learning method in the cases with and without clustering (blue lines), the expected SARSA-based method (dashed line with square shapes) exhibited a lower MAE value, even when clustering was not applied in Figure 6(a). The results indicated that the expected SARSA-based method was more efficient, in terms of managing the WPG forecast uncertainty by the ESS operation, than the Q-learning-based



(a) Mean absolute error



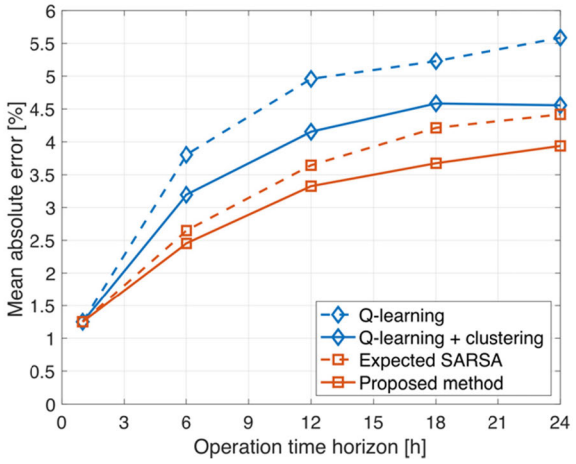
(b) Optimality gap

FIGURE 5. Mean absolute error (MAE) comparison between the Q-learning, proposed, and optimal methods. The proposed method exhibited an optimality gap of less than 0.8%-point optimality gap.

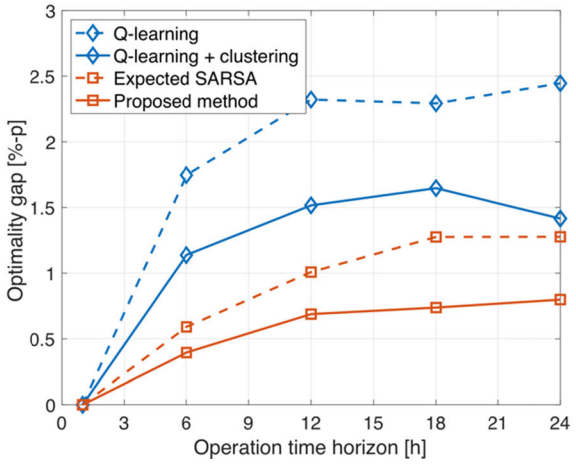
method. The difference between the results presented by the dashed and solid lines indicates that MAE improvement when clustering was applied. As shown in Figure 6(b), MAE improvement is much higher in the Q-learning-based method than the expected SARSA-based method. This indicates that the Q-leaning-based method significantly depends on the variance of the data, implying that the expected SARSA-based method is more suitable for ESS operations related to WPG with significant forecasting error variance. However, only the proposed method yielded an optimality gap within 1% of the optimal result, which requires complete information, including future values.

D. EFFECT OF CLUSTER

Figure 7 presents the changes in the MAE with respect to the number of clusters when using the proposed method. In most cases, with increasing number of clusters, the MAE was found to improve. The results indicate that clustering effectively reduces the variance of the WPG forecast uncertainty,

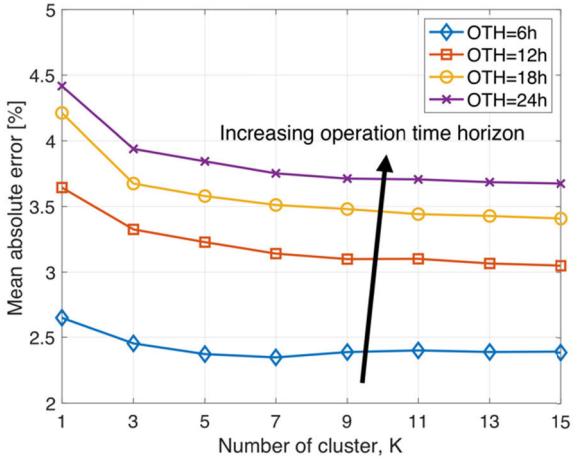


(a) MAE

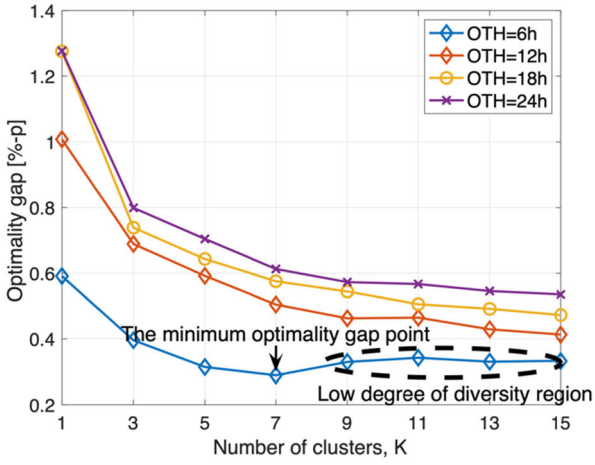


(b) Optimality gap

FIGURE 6. MAE comparison with respect to a combination of methodologies. Only the proposed method yielded an optimality gap within 1% of the optimal result.



(a) MAE



(b) Optimality gap

FIGURE 7. Changes in MAE with respect to the number of clusters K. The results indicate that clustering effectively reduces the variance of the WPG forecast uncertainty, thus improving the learning performance. However, in the case, the length of the data sequence was 6, the data was not appropriately classified owing to a low degree of diversity with increasing number of clusters.

thus improving the learning performance. When the ESS operation time horizon was 12 h, 18 h, and 24 h, the MAE exhibited a similar slope in accordance with an increase in the number of clusters, as shown in Figure 7(a). This indicates that the data sequences were sufficiently diverse for clustering to be performed in these cases. However, with an operation time horizon of 6 h, the MAE exhibited the minimum optimality gap in the case wherein seven clusters were employed, as shown in Figure 7(b). In this case, the length of the data sequence was 6, which was the same as the operation time horizon; and the variance of the data was slight during this time period, as shown in Figure 4. With increasing number of clusters, the data was not appropriately classified owing to a low degree of diversity, as indicated by the black-dashed ellipsis in Figure 7(b).

Figure 8 presents the effectiveness of clustering, which is based on the improvement in the MAE in accordance with an increase in the number of clusters related to the MAE without clustering. In the figure, the case with an operation time horizon of 18 h was found to be the most effective. This indicates that the WPG forecasting has the highest degree of diversity at an operation time horizon of 18 h, instead of 24 h. Moreover, with increasing number of clusters, the effectiveness decreased because of the decreased diversity margin. However, the amount of memory required to implement the proposed method in the system linearly increased in accordance with the increasing number of clusters. Therefore, the number of clusters should be determined considering the target MAE performance in addition to the system memory and the diversity of the data sequence.

E. EFFECT OF SIZE

Figure 9 presents changes in the MAE with respect to the ESS size when the proposed method was applied with three

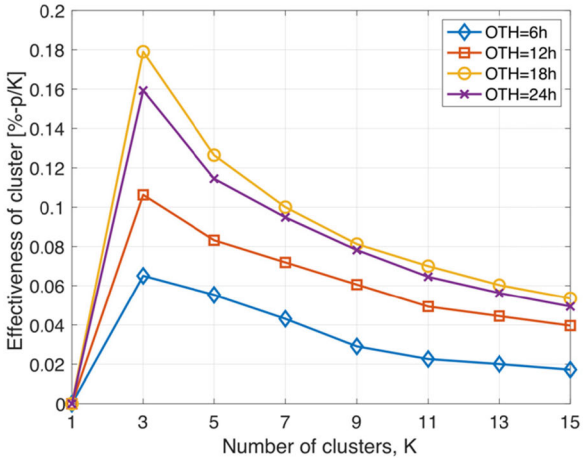
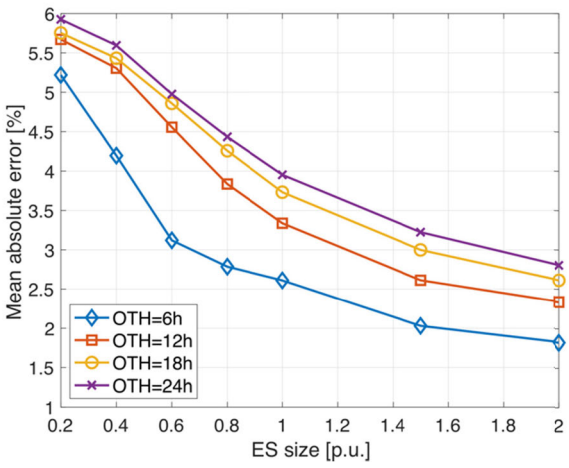


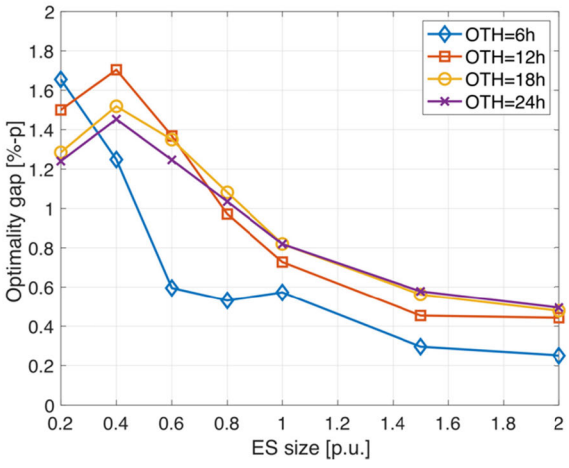
FIGURE 8. Effectiveness of clustering. The case with an operation time horizon of 18 h was found to be the most effective. This indicates that the WPG forecasting has the highest degree of diversity at an operation time horizon of 18 h, instead of 24 h in this case.

clusters. The results were obtained by changing the ESS size at a fixed C-rate of 2. The MAE improved in accordance with the increasing ESS size, as shown in Figure 9(a). This is because the ESS operation available range increased in accordance with the increasing ESS size. Similarly, the optimality decreased in accordance with increasing ESS size, as shown in Figure 9(b). This indicates that the proposed method efficiently manages the WPG forecast uncertainty, given a sufficient ESS operation size. For a small ESS size, the proposed method exhibited poor performance, as indicated by the results shown in Figures 9(a) and 9(b). Particularly, the optimality gap with 0.4 p.u. ESS size performed worse than that with 0.2 p.u. ESS size when the ESS operation time horizon was 12 h, 18 h, and 24 h in Figure 9(b). The optimal method with the small ESS size effectively managed the WPG uncertainty because it perfectly predicted future values. However, the RL-based method that is a model-free approach has less operation gain with a small operational budget increment (i.e., small ESS size). Further, the uncertainty variance affects the RL-based. Therefore, when the uncertainty variance was small such as a 6-h ESS operation time horizon, the optimality gap reduced with increasing ESS size.

Figure 10 presents the effectiveness of ESS size based on the improvement in the MAE with respect to the MAE without ESS. As discussed above, in the cases wherein a small-sized ESS was employed, a low effectiveness was observed owing to the limitations of the model-free method. The effectiveness was found to have a maximum value of approximately 1 p.u. when the operation time exceeded 12 h, and the maximum point moved to 0.6 p.u. when the operation time horizon was 6 h. This is because the WPG forecast uncertainty variance was slight in the case with an operation time horizon of 6 h. Moreover, the effectiveness decreased in the cases wherein large-sized ESSs were employed, irrespective of low MAE values and small optimality gaps.



(a) MAE



(b) Optimality gap

FIGURE 9. Changes in MAE with respect to ESS size. The results show that the proposed method efficiently manages the WPG forecast uncertainty, given a sufficient ESS operation size, but the model-free approach based proposed method has less operation gain with a small operational budget increment.

The ESS size determines the ESS installation cost; thus, it should be determined under the consideration of the target MAE performance, and the effectiveness should be determined with respect to the ESS size.

F. SUMMARY OF EFFECTIVENESS

With more resources, i.e., an increased number of clusters and ESS size, highly significant MAE improvement can be achieved. However, this increases the implementation cost. Therefore, the proposed method should be implemented effectively. Moreover, the data presented in Figures 8 and 10 can be considered a critical indicator of effectiveness. Figure 11 presents the expected MAE improvement with the application of the factor, and an operation time horizon of 24 h. The effectiveness of the ESS sizing was found to be greater than that of the clustering; thus, it was concluded that the MAE improvement depends more

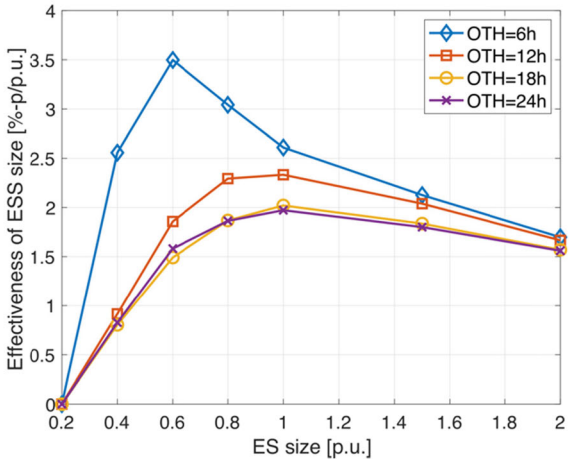


FIGURE 10. Effectiveness of ESS sizing. In the cases wherein a small-sized ESS was employed, a low effectiveness was observed owing to the limitations of the model-free method.

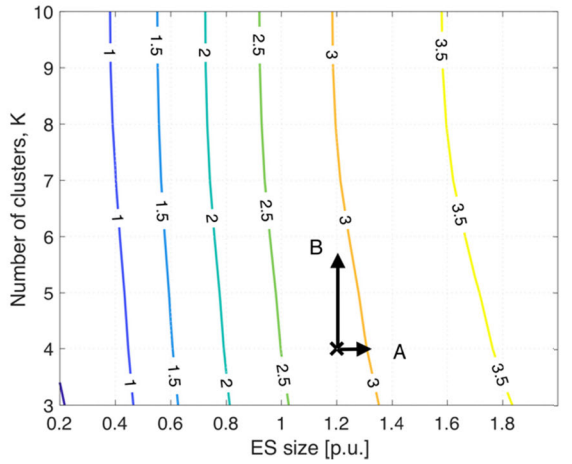


FIGURE 11. Expected MAE improvement with respect to the number of clusters and ESS size, with an operation time horizon of 24 h. For an improvement of the MAE from the point indicated by the "x" symbol, A direction is to increase the ESS size, and B direction is to increase the number of clusters.

significantly on ESS size. However, considering the memory cost required for an increase in the number of clusters in addition to the ESS capacity cost for ESS sizing, an increase in the ESS size may not be a cost-effective method. For example, with reference to Figure 9; for an improvement of the MAE from the point indicated by the "x" symbol to 3, the operator can increase the ESS size (A direction) or increase the number of clusters (B direction). If the cost of increasing four clusters is lower than the cost of growing the ESS size by 0.2 p.u., then increasing the cluster is a cost-effective way. Consequently, the effectiveness of resources can be used to determine system implementation with the target performance.

G. SUMMARY OF RESULTS

All results are summarized in Table 1 and 2. Table 1 shows performance summary according to methods. The Q-learning-based method has the lowest MAE improvement,

TABLE 1. Results according to methods.

Methodology	MAE improvement
Q-learning	Low
Q-learning + clustering	Medium
Expected SARSA	Medium, but more Q-learning + clustering
Proposed method (Expected SARSA + clustering)	High

TABLE 2. Results according to resources.

Resource	MAE improvement	Cost
Clustering	Medium	Medium (Memory & computation cost)
ESS size	High	High (ESS cost)

and the clustering enhances the MAE improvement of the Q-learning-based method. However, the expected SARSA-based method shows the more MAE improvement. The proposed expected SARSA-based method with clustering has the highest MAE improvement. Table 2 shows results summary according to resources such as clustering and ESS size. The ESS size increment enhances the MAE improvement more effectively than the clustering. However, the ESS is more expensive resource than the memory and computation cost for the clustering.

V. CONCLUSION

This study has focused on the RL-based ESS operation strategy for WPG forecast uncertainty management. First, the ESS operation problem has been presented as the MDP model. The state–action space of the model has been composed considering the ES and PS constraints of the ESS. The optimal leaning policy based on the MDP model has been suggested to solve the MAE minimization problem. The expected SARSA-based methods have been applied for learning because that method is more robust toward the uncertainty environment compared to the conventional Q-learning-based method. Furthermore, k-mean clustering has been combined as the frequency-domain data preprocessing. It has improved the effectiveness of the proposed RL-based strategy by reducing the WPG forecast uncertainty variance. The empirical study using the actual WPG generation and its forecasting data have indicated that the proposed strategy has yielded less than a 1%-point gap for managing the MAE to the optimal solution, which requires complete information, including future values. In addition, the effects of various parameters such as the number of clusters and ESS size have been discussed. Results have shown that the expected SARSA-based method improved the MAE about 1.5%-point compared to

the Q-learning-based method and the MAE performance has additionally improved about 0.5%-point by combining the clustering. By increasing the number of clusters, the MAE enhancement has been converged, but this MAE enhancement has continuously reduced with increasing ESS size. However, in terms of effectiveness, three clusters and 0.4 p.u. ESS size are the best effectiveness points. Utilizing the effectiveness of resources, this study serves as a basis for the implementation of the proposed method.

In future works, this study will be extended to various perspectives. From a technical perspective, the proposed method has been used as an expected SARSA algorithm. Considering the deep learning model, a deep Q-learning algorithm can be applied. Moreover, the results have shown that data preprocessing significantly impacts performance. Therefore, various data preprocessing methods such as singular value decomposition can be considered. From the system perspective, this work only considers a WPG system model. The system model can be practically extended further considering grid environments. In this model, the problem can be formulated by including addition units such as thermal generation units and demands and grid parameters such as power flow constraints, and be implemented in IEEE standard bus systems.

REFERENCES

[1] B. Dudley, "BP statistical review of world energy 2019," BP, London, U.K., Tech. Rep. #68, Jun. 2019.

[2] A. Z. Amin, "Global energy transformation: A roadmap to 2050," Int. Renew. Energy Agency, Abu Dhabi, United Arab Emirates, Tech. Rep., Apr. 2019.

[3] A. Z. Amin, "Renewable energy statistics 2019," Int. Renew. Energy Agency, Abu Dhabi, United Arab Emirates, Tech. Rep., Jul. 2019.

[4] Z. Conka, M. Kolcun, and G. Morva, "Impact of renewable energy sources on power system stability," *Power Electr. Eng.*, vol. 32, pp. 25–28, Nov. 2014.

[5] S. Oh, H. Shin, H. Cho, and B. Lee, "Transient impact analysis of high renewable energy sources penetration according to the future Korean power grid scenario," *Sustainability*, vol. 10, no. 11, pp. 1–15, Nov. 2018.

[6] Y. Wang, Q. Hu, D. Srinivasan, and Z. Wang, "Wind power curve modeling and wind power forecasting with inconsistent data," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 16–25, Jan. 2019.

[7] Y. Lin, M. Yang, C. Wan, J. Wang, and Y. Song, "A multi-model combination approach for probabilistic wind power forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 1, pp. 226–237, Jan. 2019.

[8] Y. Zhao, L. Ye, P. Pinson, Y. Tang, and P. Lu, "Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5029–5040, Sep. 2018.

[9] A. Sharifian, M. J. Ghadi, S. Ghavidel, L. Li, and J. Zhang, "A new method based on type-2 fuzzy neural network for accurate wind power forecasting under uncertain data," *Renew. Energy*, vol. 120, pp. 220–230, May 2018.

[10] H. Zhao, Q. Wu, S. Hu, H. Xu, and C. N. Rasmussen, "Review of energy storage system for wind power integration support," *Appl. Energy*, vol. 137, pp. 545–553, Jan. 2015.

[11] Y. Yang, S. Bremner, C. Menictas, and M. Kay, "Battery energy storage system size determination in renewable energy systems: A review," *Renew. Sustain. Energy Rev.*, vol. 91, pp. 109–125, Aug. 2018.

[12] A. Nikoobakht, J. Aghaei, M. Shafie-khah, and J. P. S. Catalao, "Allocation of fast-acting energy storage systems in transmission grids with high renewable generation," *IEEE Trans. Sustain. Energy*, early access, Aug. 2019.

[13] A. Nikoobakht, J. Aghaei, M. Shafie-Khah, and J. P. S. Catalao, "Assessing increased flexibility of energy storage and demand response to accommodate a high penetration of renewable energy sources," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 659–669, Apr. 2019.

[14] J. Aghaei, A. Nikoobakht, M. Mardaneh, M. Shafie-Khah, and J. P. Catalão, "Transmission switching, demand response and energy storage systems in an innovative integrated scheme for managing the uncertainty of wind power generation," *Int. J. Electr. Power Energy Syst.*, vol. 98, pp. 72–84, Jun. 2018.

[15] R. Sebastián, "Application of a battery energy storage for frequency regulation and peak shaving in a wind diesel power system," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 3, pp. 764–770, Feb. 2016.

[16] J. Tan and Y. Zhang, "Coordinated control strategy of a battery energy storage system to support a wind power plant providing multi-timescale frequency ancillary services," *IEEE Trans. Sustain. Energy*, vol. 8, no. 3, pp. 1140–1153, Jul. 2017.

[17] K. K. Mehmood, S. U. Khan, S.-J. Lee, Z. M. Haider, M. K. Rafique, and C.-H. Kim, "Optimal sizing and allocation of battery energy storage systems with wind and solar power DGs in a distribution network for voltage regulation considering the lifespan of batteries," *IET Renew. Power Gener.*, vol. 11, no. 10, pp. 1305–1315, Aug. 2017.

[18] Y. Yuan, X. Zhang, P. Ju, Q. Li, K. Qian, and Z. Fu, "Determination of economic dispatch of wind farm-battery energy storage system using genetic algorithm," *Int. Trans. Electr. Energ. Syst.*, vol. 24, no. 2, pp. 264–280, Feb. 2014.

[19] Q. Jiang, Y. Gong, and H. Wang, "A battery energy storage system dual-layer control strategy for mitigating wind farm fluctuations," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3263–3273, Aug. 2013.

[20] V. Khare, S. Nema, and P. Baredar, "Optimisation of the hybrid renewable energy system by HOMER, PSO and CPSO for the study area," *Int. J. Sustain. Energy*, vol. 36, no. 4, pp. 326–343, Apr. 2017.

[21] Y. Sun, J. Zhong, Z. Li, W. Tian, and M. Shahidehpour, "Stochastic scheduling of battery-based energy storage transportation system with the penetration of wind power," *IEEE Trans. Sustain. Energy*, vol. 8, no. 1, pp. 135–144, Jan. 2017.

[22] E. Oh and S.-Y. Son, "Energy-storage system sizing and operation strategies based on discrete Fourier transform for reliable wind-power generation," *Renew. Energy*, vol. 116, pp. 786–794, Feb. 2018.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[24] T. Liu, X. Hu, S. E. Li, and D. Cao, "Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 4, pp. 1497–1507, Aug. 2017.

[25] R. Xiong, J. Cao, and Q. Yu, "Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle," *Appl. Energy*, vol. 211, pp. 538–548, Feb. 2018.

[26] Z. Wen, D. O'Neill, and H. Maei, "Optimal demand response using device-based reinforcement learning," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2312–2324, Sep. 2015.

[27] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Appl. Energy*, vol. 220, pp. 220–230, Jun. 2018.

[28] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Appl. Energy*, vol. 235, pp. 1072–1089, Feb. 2019.

[29] B. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck, "Battery energy management in a microgrid using batch reinforcement learning," *Energies*, vol. 10, no. 11, p. 1846, Nov. 2017.

[30] B.-G. Kim, Y. Zhang, M. Van Der Schaar, and J.-W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2187–2198, Sep. 2016.

[31] X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor, "Reinforcement learning-based microgrid energy trading with a reduced power plant schedule," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10728–10737, Dec. 2019.

[32] E. Foruzan, L.-K. Soh, and S. Asgarpoor, "Reinforcement learning approach for optimal distributed energy management in a microgrid," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5749–5758, Sep. 2018.

[33] V.-H. Bui, A. Hussain, and H.-M. Kim, "Q-learning-based operation strategy for community battery energy storage system (CBESS) in microgrid system," *Energies*, vol. 12, no. 9, p. 1789, May 2019.

[34] V.-H. Bui, A. Hussain, and H.-M. Kim, "Double deep Q -learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 457–469, Jan. 2020.

[35] S. Kim and H. Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, no. 8, p. 2010, Aug. 2018.

[36] G. Li, Z. Yang, B. Li, and H. Bi, "Power allocation smoothing strategy for hybrid energy storage system based on Markov decision process," *Appl. Energy*, vol. 241, pp. 152–163, May 2019.
[37] T. Chen and W. Su, "Local energy trading behavior modeling with deep reinforcement learning," *IEEE Access*, vol. 6, pp. 62806–62814, 2018.
[38] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.
[39] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, "Deep reinforcement learning for joint bidding and pricing of load serving entity," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6366–6375, Nov. 2019.
[40] R. Rocchetta, L. Bellani, M. Compare, E. Zio, and E. Patelli, "A reinforcement learning framework for optimal operation and maintenance of power grids," *Appl. Energy*, vol. 241, pp. 291–301, May 2019.
[41] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.
[42] S. Yang, Y. Gao, B. An, H. Wang, and X. Chen, "Efficient average reward reinforcement learning using constant shifting values," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016.
[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[44] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*. Upper Saddle River, NJ, USA: Prentice-Hall, 1997.
[45] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.
[46] D. E. Kirk, *Optimal Control Theory: An Introduction*. Upper Saddle River, NJ, USA: Prentice-Hall, 1970.
[47] P. M. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012.
[48] J. Mur-Amada and A. Bayod-Rujula, *Variability of Wind and Wind Power*. London, U.K.: IntechOpen, 2010.
[49] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[50] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is NP-hard," in *Proc. Int. Workshop Algorithms Comput.* Kolkata, India: Springer, Feb. 2009, pp. 274–285.
[51] *BPA Facts*, BPA Headquarters, Bonneville Power Admin., Portland, OR, USA, Apr. 2019.

EUNSUNG OH (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Yonsei University, Seoul, South Korea, in 2003, 2006, and 2009, respectively. From 2009 to 2011, he was a Postdoctoral Researcher with the Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California. From 2011 to 2012, he was a Senior Researcher with the Korea Institute of Energy Technology Evaluation and Planning, South Korea. From 2012 to 2013, he was a Research Professor with the Department of Electrical Engineering, Konkuk University, South Korea. He is currently an Associate Professor with the Department of Electrical and Electronic Engineering, Hanseo University, South Korea. His main research interests include the design and analysis of algorithms for green communication networks and smart grids.

HANHO WANG (Member, IEEE) received the B.S.E.E. and Ph.D. degrees from Yonsei University, in 2004 and 2010, respectively. He was a Patent Examiner with the Department of Information and Telecommunication Patent Examination, Korean Intellectual Property Office. In 2012, he joined the Information and Telecommunication Engineering Department, Sangmyung University, where he currently serves as an Associate Professor.