

Article

Bidding a Battery on Electricity Markets and Minimizing Battery Aging Costs: A Reinforcement Learning Approach

Harri Aaltonen ^{1,*} , Seppo Sierla ¹ , Ville Kyrki ¹, Mahdi Pourakbari-Kasmaei ¹  and Valeriy Vyatkin ^{1,2}

¹ Department of Electrical Engineering and Automation, School of Electrical Engineering, Aalto University, FI-00076 Espoo, Finland; seppo.sierla@aalto.fi (S.S.); ville.kyrki@aalto.fi (V.K.); mahdi.pourakbari@aalto.fi (M.P.-K.); valeriy.vyatkin@aalto.fi (V.V.)

² Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, 97187 Luleå, Sweden

* Correspondence: harri.aaltonen@aalto.fi

Abstract: Battery storage is emerging as a key component of intelligent green electricity systems. The battery is monetized through market participation, which usually involves bidding. Bidding is a multi-objective optimization problem, involving targets such as maximizing market compensation and minimizing penalties for failing to provide the service and costs for battery aging. In this article, battery participation is investigated on primary frequency reserve markets. Reinforcement learning is applied for the optimization. In previous research, only simplified formulations of battery aging have been used in the reinforcement learning formulation, so it is unclear how the optimizer would perform with a real battery. In this article, a physics-based battery aging model is used to assess the aging. The contribution of this article is a methodology involving a realistic battery simulation to assess the performance of the trained RL agent with respect to battery aging in order to inform the selection of the weighting of the aging term in the RL reward formula. The RL agent performs day-ahead bidding on the Finnish Frequency Containment Reserves for Normal Operation market, with the objective of maximizing market compensation, minimizing market penalties and minimizing aging costs.

Keywords: battery storage; reinforcement learning; machine learning; primary frequency reserve; frequency containment reserve; simulation



Citation: Aaltonen, H.; Sierla, S.; Kyrki, V.; Pourakbari-Kasmaei, M.; Vyatkin, V. Bidding a Battery on Electricity Markets and Minimizing Battery Aging Costs: A Reinforcement Learning Approach. *Energies* **2022**, *15*, 4960. <https://doi.org/10.3390/en15144960>

Academic Editors: Tek Tjing Lie and Daniel T. Hallinan, Jr.

Received: 17 May 2022

Accepted: 4 July 2022

Published: 6 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Battery storage is emerging as a key component of intelligent green electricity systems [1]. The investment into the battery needs to be justified by using the battery to provide services of financial value [2,3]. Four categories of such services can be identified: The first category is energy arbitrage, in which participants store energy in the battery during low prices and sell during high prices [4]. The key electricity markets for energy arbitrage are the day-ahead, intraday, balancing, and real-time markets. Energy arbitrage can involve a standalone battery or a battery paired with other energy resources such as photovoltaic (PV) generation [5]. The second category of services provided by batteries is reserves. Reserve market participants are compensated for adjusting power consumption or generation in response to power grid imbalances, such as grid frequency deviations from a nominal value. Reserve markets have different names and specifications in different parts of the world. In the literature, some commonly used names for these markets are frequency reserves [6], frequency regulation [7], ancillary services [8] and spinning reserves [9]. The third category is local markets such as peer-to-peer trading [10] and the recently emerging nodal markets, in which prices vary according to the geographic location of the energy producing or consuming units, which gives the market tools to avoid congestion in the power grid [11,12]. The fourth category of services involves coordinating a battery and controllable loads [13] for the provision of demand response.

The above-mentioned services are monetized through market participation, which usually involves bidding. Bidding is a multi-objective optimization problem, involving targets such as maximizing market compensation and minimizing penalties for failing to provide the service and costs for battery aging. A subset of research in this field includes the minimization of battery aging costs by embedding those into the multi-objective optimization problem (e.g., [14,15]). However, the optimization problem is challenging to solve, because the aging phenomenon is non-linear [16] and dependent on the type of service provided, as some services, such as fast regulation services, may involve frequent charge–discharge cycling [17].

Recently, numerous researchers have proposed reinforcement learning (RL) as a multi-objective optimization technique for monetizing battery storage, and several of them consider aging in the reward formulation [18,19]. The RL problem formulation involves a reward, which gives the RL agent feedback about how advantageous or disadvantageous its actions have been. The reward does not need to be derived from physics, so the reward formula may include a term that is a simplification of the aging phenomena that ignores the non-linear dynamics of the battery. Such simplifications are commonly used by RL practitioners, and various formulations have been proposed by different authors, such as [20]. The shortcoming of such an approach is that any benefits for reducing battery aging are not demonstrated either with a physical battery or with an accurate battery model. This prevents researchers from validating effective formulations of battery aging in the RL reward function. Simplified aging cost models also prevent making direct comparisons between results reported by different research groups, preventing the identification of seminal works with respect to battery aging management. **The contribution of this article is a methodology involving a realistic battery simulation to assess the performance of the trained RL agent with respect to battery aging in order to inform the selection of the weighting of the aging term in the RL reward formula.** Our research is conducted in the context of bidding battery storage on frequency reserves. A few works have investigated using RL for bidding a battery on frequency reserves, but without considering battery aging effects [21,22] or through simple approaches such as penalizing the agent for exceeding minimum and maximum state of charge (SoC) limits for the battery [23].

2. Literature Review

2.1. Financial Exploitation of Battery Storage

In Section 1, energy arbitrage, reserves, local markets, and demand response were introduced as general categories of battery services. Examples of applications to actual electricity markets are as follows. Solutions for prosumer buildings with PV generation and battery storage include energy arbitrage in the day-ahead Croatian Power Exchange market [24] and the Electric Reliability Council of Texas (ERCOT) nodal market [25]. Pre-studies to support sizing decisions when purchasing the battery are performed to estimate the revenue of a large-scale battery storage participating in the Finnish Frequency Containment Reserves for Normal Operation market [26], the South China Region frequency regulation market [27], and the real-time market in Queensland, Australia [28]. In the case of a nodal market, the pre-study should optimize the siting and sizing of a battery, and such a solution is demonstrated for the New Zealand nodal wholesale market [29]. Energy arbitrage solutions for standalone batteries are proposed for the Spanish real-time electricity market [30], Italian balancing market [31] and the California Independent System Operator (CAISO) day-ahead market [32].

Other authors report increased financial benefits from bidding on more than one market. Such market combinations include frequency regulation and energy arbitrage [33], frequency regulation and real-time markets [34], energy arbitrage and reserve markets [35], energy arbitrage and spinning reserves [36], energy arbitrage and demand response [37] and demand response and nodal markets [38]. It is notable that some authors specify their targeted market in general terms, such as energy arbitrage, while others identify a more specific market, such as real-time electricity markets.

The emergence of Vehicle-to-Grid (V2G) solutions enables the provision of similar services with electric vehicle batteries when the vehicle is plugged in, if permitted by the vehicle owner. V2G applications are complicated by the fact that vehicle users have requirements on the battery SoC at the time when they intend to use the vehicle. This will increase the demand for fast charging, in which case the thermal safety needs to be considered by the RL agent [39]. One approach is to define penalty component to the reward function in case temperature exceeds a threshold [40]. Further, V2G needs to cope with the vehicle being unplugged unexpectedly in violation of a contract made with the vehicle owner [41]. Despite these challenges, V2G has been used to exploit electric vehicle batteries for ancillary services [42], demand response [43,44], local markets [45] and frequency regulation [46]. In the absence of V2G capabilities, an electric vehicle charger may perform frequency regulation by curtailing or rescheduling the charging [47,48].

2.2. Reinforcement Learning Applications for Batteries

De Silva et al. [49] propose a machine learning architecture for exploiting distributed energy resources in various energy markets. Forecasts of market prices and energy resource availability are fed into a bidding optimizer. Such forecasts based on machine learning time series forecasts are readily available for markets such as real-time markets [50], frequency reserve markets [51] and day-ahead spot markets [52], as well as for the availability of resources such as PV generation [53], wind generation [54], and building energy consumption [55]. RL could emerge as the ideal technology for the bidding optimizer exploiting such forecasts. Currently, RL has been demonstrated as a potential multi-objective optimization technique for performing services such as the ones discussed in Section 2.1. The majority of the research investigates energy arbitrage and local markets. Only a minority of these works consider standalone batteries [56], whereas the rest optimize the battery alongside other distributed energy resources. In the context of a building with local PV generation, stationary batteries [57,58], one-way electric vehicle charging [59,60] and V2G [56] batteries have been used for energy arbitrage. Another form of energy arbitrage is achieved through pairing batteries and wind power generation [61]. Similar solutions are applicable to micro-grids with local markets [62,63]. Only a few works consider frequency reserves [21–23] or ancillary services [64].

Battery aging is considered in some of the works applying RL to batteries. One approach is to define maximum and minimum permissible values for the SoC and charging and discharging currents as constraints, which are enforced outside the RL problem formulation. For example, if the RL agent just turns the batteries on or off, a separate algorithm can determine the charging and discharging currents [65]. Another approach is to define a logic that overrides the action of the RL agent if these constraints are violated [66]. However, it is also possible to incorporate battery aging penalties into the RL reward function, in which case the agent can learn to mitigate aging. If the penalty is defined in terms of SoC, the following approaches are encountered in the literature: exceeding minimum or maximum SoC limits [18], being outside of an ideal SoC range [67], as a deviation from a reference SoC value [68], or in terms of depth of discharge [69]. Other authors define the aging penalty in terms of the magnitude of charging and discharging power [70,71]. Some authors have added an overheating component to the aging penalty. This can be defined as a temperature deviation above a maximum temperature value [72] or as the second exponent of such deviation [73].

Each of these approaches can be expected to direct the agent to operate the battery in a range that is advantageous for mitigating aging. However, the actual impact on battery aging is not validated on physical batteries or with accurate battery simulation models. Validation on physical batteries is impractical, as there is no easy way to measure aging and since the experiments could be very time-consuming. Thus, in this article, an aging penalty will be defined for the reward function building on the above-mentioned approaches, and additionally, a validation of aging effects will be performed with a simulated battery.

3. Methodology

The methodology is elaborated in this section. As frequency reserve market rules vary between countries, the Finnish Frequency Containment Reserve for Normal operation (FCR-N) is taken as the case market. The symbols used to define the methodology are presented in Table 1.

Table 1. Symbols.

Symbol	Description
SoC	State of charge of the battery in percent
DoD	Depth of discharge—the complement of SoC , i.e., $DoD = 100\% - SoC$
i	Charging/discharging current of the battery—negative values mean discharging
u	Voltage of the battery
P	Charging/discharging power of the battery, $P = u \cdot i$
R	Hours since the battery last rested, i.e., did not participate on the market and was charged/discharged to 50% SoC with constant power
FCR_{act}	Actual price (EUR/MW) on that particular hour on the Finnish FCR-N market
FCR_{fcast}	Forecasted price (EUR/MW) on that particular hour on the Finnish FCR-N market
T	Number of seconds in one market interval of FCR-N. At the time of writing, $T = 3600$
C	Capacity of the bid made to the FCR-N market. The capacity is in MW and specifies the value of P when the full activation is required. The full activation occurs when the power grid frequency deviation exceeds the limit at which the FCR-N technical specification requires 100% activation of the capacity. In our RL formulation, four discrete values for capacity are used: 0 MW, 0.3 MW, 0.4 MW and 0.5 MW
age	Age of the battery (equivalent full cycles)
pen	Market penalty for failing to deliver the capacity C . The penalty is expressed as a percentage of the duration of the market interval, during which the battery is not available due to the SoC exceeding a minimum or maximum threshold
A	Aging penalty (component of the reward function)
w	Weight for the aging penalty A

We propose to model the reward for the FCR-N market using three components: market revenue, market penalty, and aging penalty. The reward can then be defined as:

$$\text{reward} = \frac{100 - pen}{100} \cdot FCR_{act} \cdot C - \frac{pen}{100} \cdot FCR_{act} \cdot C - wA \quad (1)$$

where the first two terms are the FCR-N market revenue and market penalty, respectively [21]. The market penalty is due to the battery SoC exceeding minimum or maximum limits, in which case the battery is not available for reacting to grid frequency deviations. The third term is an aging penalty A and a weight w . The value of w determines the weight of the aging penalty relative to the net revenues in the reward function. It is notable that the market prices in one hour can be much higher than in another hour, so a high reward can be due to good decisions by the agent or due to high prices on that particular hour or both.

For the aging penalty in reward, we model aging as a linear approximation of the non-linear battery dynamics. The aging can be modelled as:

$$A = \frac{1}{10^7} \sum_{s=0}^{T-1} i_s \cdot DoD_s \quad (2)$$

where the coefficient $\frac{1}{10^7}$ scales the sum to same level with other components of reward.

The approximation in (2) builds on top of the previous research reviewed in Section 2.2, but introduces two key differences. Firstly, the works discussed in Section 2.2 identify either SoC or i as significant factors impacting aging. Our formulation recognizes that aging depends on the SoC level as well as the magnitude of the charging/discharging current. Secondly, the step of our RL agent is the market interval, which for FCR-N is 1 h at the time of writing. Power grid frequency can change numerous times during this interval, resulting in a corresponding change in the current i , which will impact SoC and DoD . Thus, a much more accurate approximation of aging can be obtained by performing the approximation once per second and taking the sum over the market interval, which is the duration of the RL step. The reason for performing the approximation once per second is that a control step of one second is sufficient for meeting the dynamic and stability requirements of the FCR-N market [74].

Figure 1 presents an architecture for training the RL agent. Since the step of the RL agent is 1 h, the state and reward are updated once per hour, and the RL agent determines the bid capacity C once per hour. However, the calculation of pen in (1) requires a once per minute resolution [21], and the calculation of A in (2) requires a once per second resolution. Thus, the environment requires simulation at a 1st time step. As presented in Figure 1, this is driven by time series data of the power grid frequency f , which has been obtained from the transmission system operator (TSO) Fingrid, which also operates the FCR-N market. The frequency data have been preprocessed to obtain one data point per second. A ‘Power calculator’ module determines the required momentary charging/discharging power P based on f , C and the stationary requirement of FCR-N [74,75]. The battery simulation determines the required current i according to P and the u , which is not assumed to be constant as it is affected by the SoC . This current is fed as an input to a battery simulation model, which outputs the SoC and i . This information is sufficient for calculating according to (1) and (2). These calculations are done based on the actual market price FCR_{act} . However, this is not known at the time of making the bid, so it cannot be used as state information for the RL agent. Thus, the state includes the forecasted price $FCR_{forecast}$, which is obtained using the machine learning time series forecasting method for FCR-N [76]. In addition to this forecast, the state information includes R , which is an integer specifying the number of hours since the battery last rested. Resting is defined as not participating in the market, which occurs when the bid capacity C is 0 MW. During the rest, the battery is charged or discharged so that the SoC will reach 50%, reducing the likelihood of SoC out of bounds events that result in market penalties.

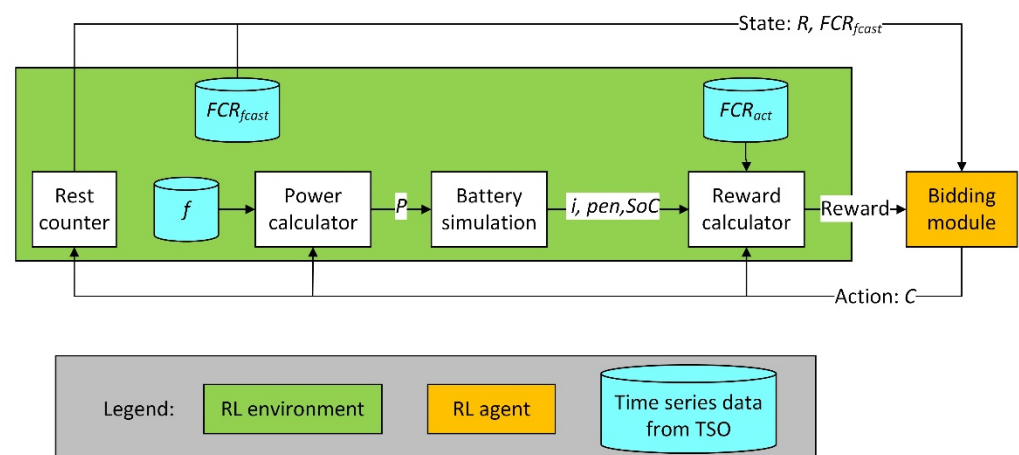


Figure 1. Architecture for training the RL agent.

Several RL algorithms are available for optimizing the agent. Many of the RL applications for battery management define continuous action spaces, which motivates the selection of algorithms capable of handling continuous as well as discrete spaces, such as Advantage Actor-Critic (A2C) [77], Proximal Policy Optimization (PPO) [22], Deep

Deterministic Policy Gradient (DDPG) [78,79] and Twin Delayed DDPG (TD3) [80] have been applied in the context of batteries. However, the task of the RL agent in this paper is to select the value for C . This selection must be made from a discrete set of possible values, due to the rules of the FCR-N market. The range of bids is between 0.1 MW and 5 MW with a resolution of 0.1 MW [21]. Since our state and action spaces are discrete and not large, computationally heavy methods such as DDPG and TD3 will not be investigated. In this article, the suitability of the REINFORCE [81], A2C [81], and PPO algorithms will be experimentally evaluated.

Equation (2) is proposed as a reasonable approximation of battery aging cost for the purpose of training the RL agent. As it does not capture a battery's non-linear dynamics, it cannot be used for an accurate evaluation of how the trained RL agent mitigates battery aging. A modification of the architecture of Figure 1 will be used for performance evaluation, in which the battery *age* is obtained from a battery simulation. The modified architecture is presented in Figure 2. No reward calculation is conducted at this stage since the training process of the agent has already been completed. The setup in Figure 2 will determine the net revenue in EUR and the battery aging in terms of full equivalent cycles when the trained agent is run against historical market and grid frequency data for a period of several days. In our RL formulation, an episode is one day. Net revenue is defined as market compensation minus market penalties, and this calculation is performed in the 'Revenue calculator' of Figure 2. The 'Battery simulation' in Figure 2 includes the Matlab Simulink battery model, which implements the aging behavior modeled in [82].

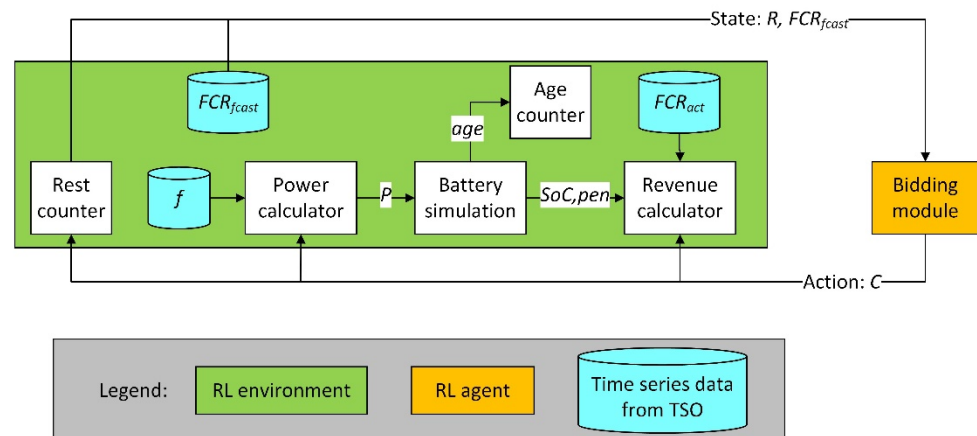


Figure 2. Architecture for performance evaluation of the trained RL agent.

In the training phase (Figure 1), Equation (2) is used instead of the aging output of the Matlab Simulink battery model (Figure 2). The reason for this is that the Simulink battery model is not intended for applications in which the *age* needs to be updated frequently, such as every time the RL environment is stepped forward. The *age* output of the Simulink battery is updated once every half cycle, and it cannot be assumed that these updates would occur at the same time as the RL environment is stepped forward.

4. Implementation

The implementations of the two architectures in Figures 1 and 2 are presented in Figures 3 and 4, respectively. In both implementations, the battery voltage is dependent on the extracted capacity [83], which affects *SoC* directly. The voltage of the battery is used as an input variable for the *CurrentCalculator* function, which generates the control signal to the controllable current source. The *CurrentCalculator* function is also responsible for preventing the *SoC* from exceeding the 5% and 95% limits. The *PowerCalculator* function is responsible for controlling the power in case of rest action (i.e., no bid). If the rest action is taken, the battery will be charged or discharged to 50% *SoC* at constant power. The blocks

between the *penaltyIn* and *penaltyOut* variables keep track of penalty minutes, which are used to calculate *pen*.

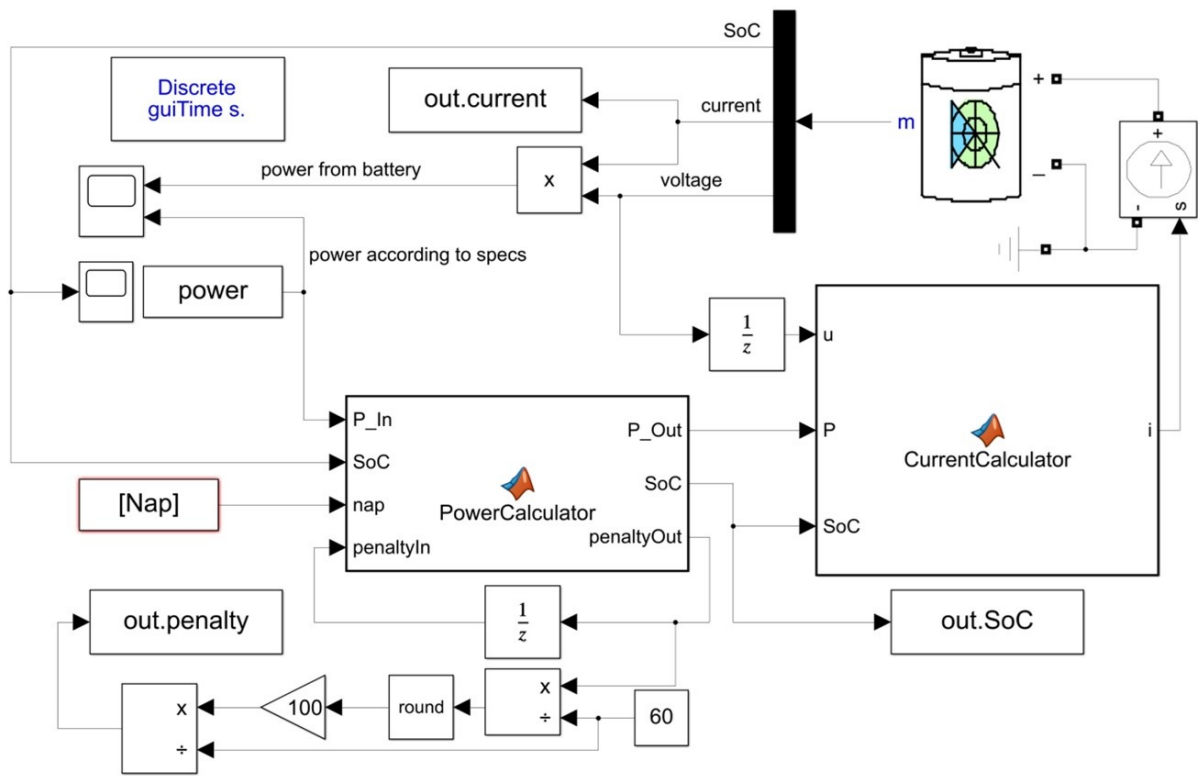


Figure 3. Simulation model for training and validation.

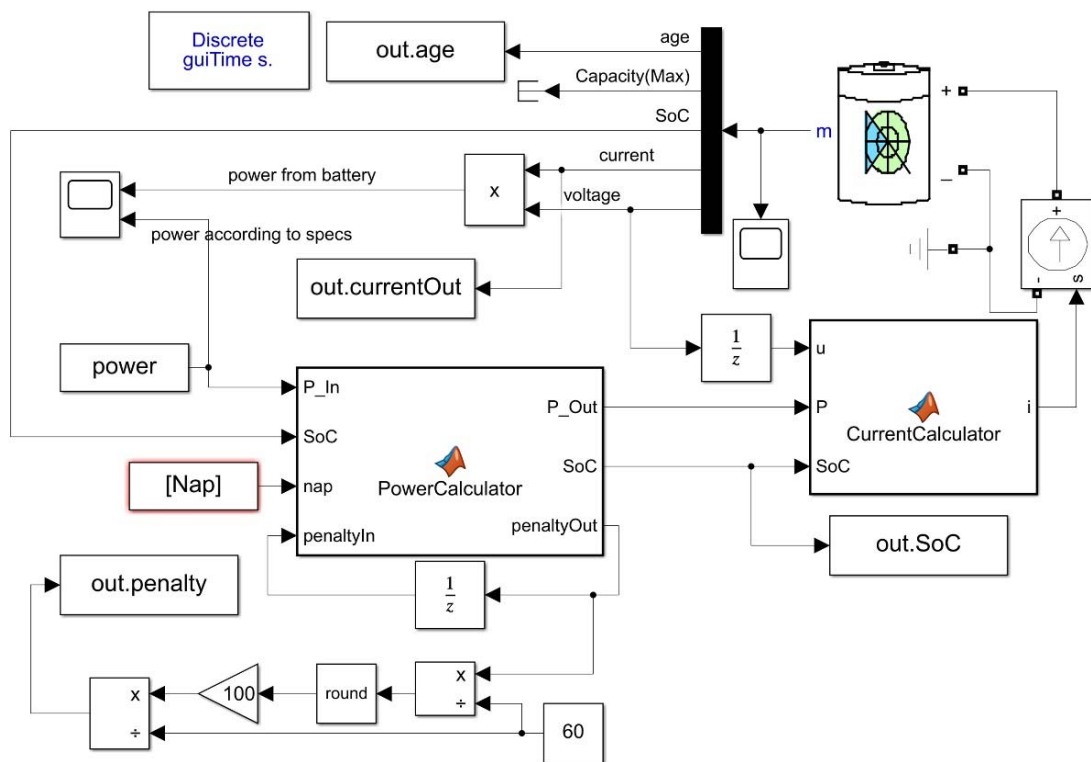


Figure 4. Simulation model for performance estimation.

The setup in Figure 3 is used to compute the reward function in (1). The aging output of the Simulink battery is not used in this context since it is updated only every half cycle, so in general, the *age* output is not up to date at the end of each RL step. For this reason, the simplified approximation of aging behavior was used as defined in (2). However, it is important to assess how well the RL agent trained with this reward will perform against the more realistic battery dynamics. The performance evaluation setup in Figure 4 is used for this purpose. The *age* output of the Simulink battery in Figure 4 is used to quantify the aging in equivalent full cycles in the performance evaluation phase. The aging dynamics of the Simulink battery model are based on the [82]. Equations (3)–(5) are from the Mathworks ‘Battery’ documentation [84]. The aging output is calculated as

$$age = CL_{100} \cdot \epsilon_n \quad (3)$$

where CL_{100} is the number of cycles when battery is fully charged and discharged at nominal charge and discharge current. The CL_{100} is an input parameter to determine how many full cycles battery lasts. The ϵ_n is the battery aging factor. The aging factor is calculated as:

$$\epsilon_n = \begin{cases} \epsilon_{n-1} + \frac{0.5}{N_{n-1}} \left(2 - \frac{DoD_{n-2} + DoD_n}{DoD_{n-1}} \right), & \text{when half-cycle update occurs} \\ \epsilon_{n-1}, & \text{otherwise} \end{cases} \quad (4)$$

The half-cycle update occurs when battery starts to discharge after charging or when the battery is full, i.e., $SoC = 100$. The DoD values are from previous three timesteps. The N_n is the maximum number of cycles and it is calculated as:

$$N_n = H \left(\frac{DoD_n}{100} \right)^\xi \cdot \exp \left(-\psi \left(\frac{1}{T_{ref}} - \frac{1}{T_{a_n}} \right) \right) \cdot \left(I_{dis\ ave_n}^{-\gamma_1} \right) \left(I_{ch\ ave_n}^{-\gamma_2} \right) \quad (5)$$

The maximum number of cycles is dependent on the average currents during latest half cycle duration, previous DoD and ambient and reference temperatures. The symbols in (5) are presented in Table 2. The constant values of N_n are set by Matlab battery model and not available from documentation.

Table 2. Symbols used in Equation (5). Adapted from [84].

Symbol	Description	Constant Value
H	The cycle number constant in cycles	YES
ξ	The exponent factor for DoD	YES
ψ	The Arrhenius rate constant for the cycle number	YES
T_{ref}	Reference temperature	YES
T_{a_n}	Ambient temperature	YES
$I_{dis\ ave_n}$	The average discharge current in A during a half cycle duration	NO
$I_{ch\ ave_n}$	The average charge current in A during a half cycle duration	NO
γ_1	The exponent factor for the discharge current	YES
γ_2	The exponent factor for the charge current	YES

The parameters of the assessed algorithms for training and validation are presented in Table 3. The parameters of the battery are presented in Table 4. The additional parameters for performance evaluation are presented in Table 5.

Table 3. Parameters for training and validation.

	REINFORCE	A2C	PPO
Gamma (discount factor)	0.5	0.5	0.5
Alpha (learning rate)	0.01	0.01	0.01
Input states	2	2	2
Number of hidden layers	1	1	1
Nodes on hidden layer 1	20	20	20
Hidden layer activation function	Sigmoid	Sigmoid	Sigmoid
Output actions	4	4	4
Output layer activation function	Softmax	Softmax	Softmax
Dropout	Not used	Not used	Not used
Optimizer	Adam	Adam	Adam
Number of seeds	10	10	10
N (length of memory)			24
Batch size			6
Number of epochs			4
Gae lambda			0.95
Policy clip			0.2

Table 4. Battery parameters.

Nominal voltage (V)	772.8
Rated capacity (Ah)	700.0
Battery response time (s)	2.0
Sample time (s)	1.0
Maximum capacity	700.0
Cut-off voltage (V)	504.0
Fully charged voltage	907.4
Nominal discharge current (A)	304.3478
Internal resistance (Ohms)	0.01104
Capacity (Ah) at nominal voltage	633.0435
Exponential zone [voltage (V), capacity (Ah)]	[834.9222, 34.3913]

Table 5. Parameters for battery aging.

Initial battery age (equivalent full cycles)	0.0
Aging model sampling time (s)	2.0
Ambient temperatur Ta1 (deg. C)	25.0
Capacity at End of Life (EOL) (Ah)	633.0435×0.9
Internal resistance at EOL (Ohms)	0.01104×1.2
Charge current (nominal, maximum) [Ic (A), Imax (A)]	[304.3478, 304.3478×3]
Discharge current (nominal, maximum) [Id (A), Idmax (A)]	[304.3478, 304.3478×3]
Cycle life at 100% DOD, Ic and Id (cycles)	20,000.0
Cycle life at 25% DOD, Ic and Id (cycles)	40,000.0
Cycle life at 100% DOD, Ic and Idmax (cycles)	8000.0
Cycle life at 100% DOD, Imax and Id (cycles)	8000.0
Ambient temperatur Ta2 (deg. C)	45.0
Cycle life at 100% DOD, Ic and Id (cycles)	950.0

For training and validation, the predicted and actual prices of the Finnish FCR-N market and the Finnish power grid frequency data from 2020 were used. One episode is one day, and one step of the RL agent is one hour since the market interval is one hour. The RL environment was reset at the beginning of each episode. The days were shuffled and then split into training and validation datasets with a ratio of 9:1. In the data preprocessing phase, any days with missing data were excluded, resulting in 315 training days and 35 validation days; 10 random seeds were used to train 10 agents for each RL algorithm.

The tunable weight term w in the reward restricts the learning of the agents. The teaching is meaningful only if the aging penalty is significant, but it does not dominate the net revenue. If the net revenues dominates the reward, the agent is expected to ignore aging penalties, and if the aging penalty dominates, then there is no business case since costs outweigh the revenues. The different components of the reward function were plotted for several values of w and it was determined that a value of 2.63 was in the meaningful range as described above. The methodology is presented in detail using this value of w . The performance evaluation is performed for several values of w in Section 5.2.

Since the state space has only two variables, the mapping from the state space to the actions learned by the RL agent can be visualized as a heatmap, in which the horizontal and vertical axes are the values of the state variables and the color is the value of the bidding action. Figure 5 shows this mapping for each of the three algorithms: REINFORCE (a), A2C (b), and PPO (c). Since 10 random seeds are used, the action values are the mean of actions selected by the 10 agents. The learned policies of the three algorithms display a triangular pattern, which can be intuitively explained. The longer the time since the previous rest, the higher the likelihood of the SoC going out of bounds and incurring market penalties, and the higher the market price, the higher the revenue. The agent learns to capitalize on this phenomenon by using higher bids toward the top right corner of the heatmap.

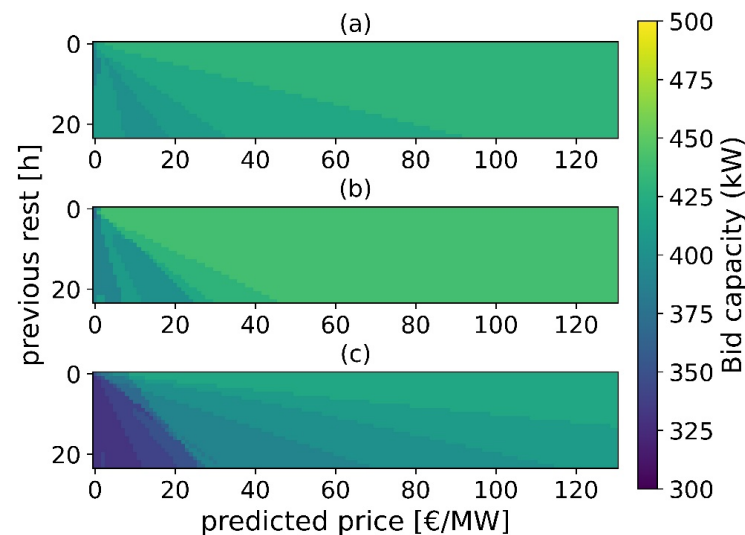


Figure 5. Action selections of three algorithms: (a) REINFORCE, (b) A2C and (c) PPO.

5. Results

The proposed method was evaluated experimentally to address the following questions: How does the selection of w impact market net revenue and battery aging, and how does the selection of the RL-algorithm impact learning?

5.1. Training and Validation

The training dataset consisting of 315 days was presented to the agents 5 times. Figure 6 shows the exponentially weighted mean with adjustment of training reward. Such filtering was needed since the rewards are dependent on the FCR_{act} , and the fluctuation, caused by the market price, is not indicative of the training performance. The weight term

$1 - \alpha$ was set to 0.995. Figure 6 has a seasonal component that repeats 5 times, which is due to the dataset being presented to the agent 5 times before terminating the training. After the agent has seen the dataset 3 times, only minor improvements are seen in the reward. Figure 7 has the same information as Figure 6, with the shaded areas being the standard deviations resulting from our use of 10 seed values. It is noteworthy to mention that the training performance of all three algorithms is very similar, and after examining the standard deviations, none of the algorithms are statistically superior.

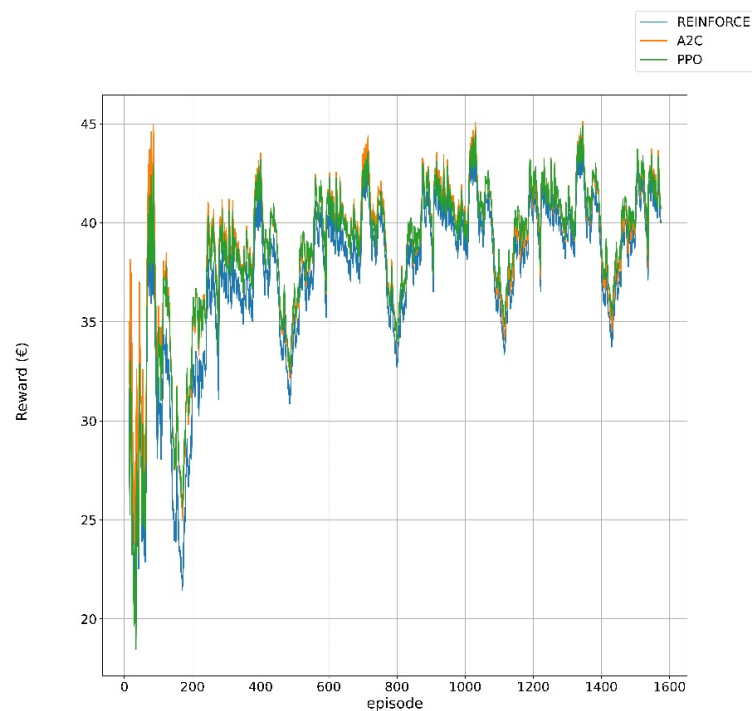


Figure 6. Training reward.

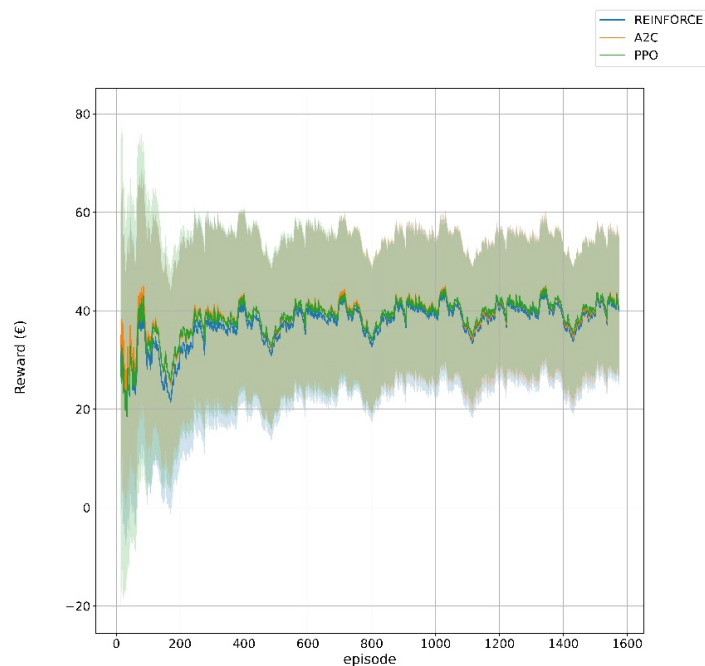


Figure 7. Training reward with standard deviation included.

Our dataset consisted of 350 days and 10% of the data were reserved for validation. Since the FCR-N market has a strong seasonality, the dataset was shuffled and 35 days

were selected for the validation dataset. The validation is done after every 15 episodes of training. The validation reward is the cumulative reward for these 35 days. The purpose of the validation is to confirm that learning occurs and to identify the episode at which the learning plateaus. The validation is repeated for 10 random seeds for each algorithm. The standard deviation of the validation reward of each algorithm is plotted to determine whether the performance of any of the algorithms is statistically superior to the others.

Towards the last episode, only very minor improvements occur in the reward. Figure 8 shows this cumulative reward for 35 days. Figure 9 shows the same as Figure 8 with standard deviations for 10 seeds included. None of the algorithms is statistically better than the others.



Figure 8. Validation reward.

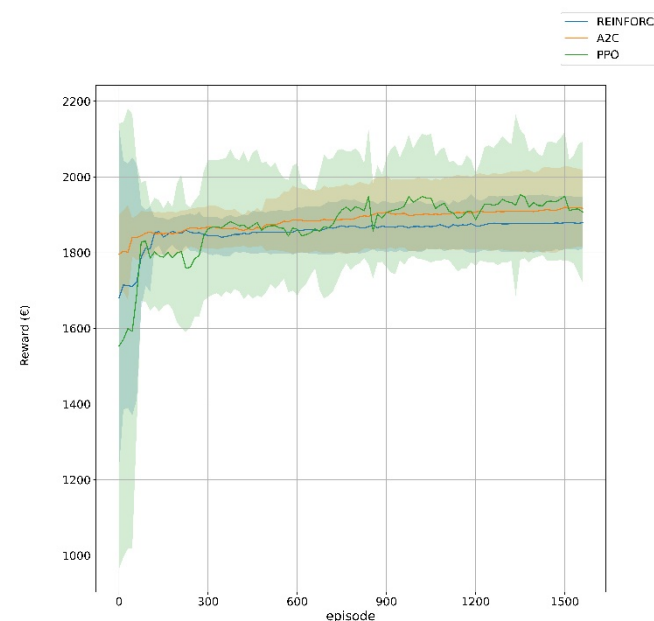


Figure 9. Validation reward with standard deviation.

5.2. Performance Evaluation

The validation rewards indicate that learning occurs with respect to our reward function. As has been explained in Section 4, a simplified model of battery aging has been used in the reward function since the battery simulation model is not intended to update the *age* output at each RL step. This approach is common in the literature. However, it is unclear how such RL agents would perform against a real battery. In this paper, our interest is to further investigate how an RL agent trained and validated on such a simplified model of battery aging would perform if connected to a more detailed model capturing the aging dynamics in detail. The setup in Figure 4 has been used for such a performance evaluation. The agents that have been trained as described in Section 5.1 are run for the full duration of the validation set (35 days). The initial value for the battery *age* is zero, and the *age* output is recorded at the end of the 35-day simulation. Since 10 seed values have been used for each of the three RL algorithms, the performance evaluation run is performed 30 times for each value of w .

Figure 10 is the scatter plot of Net revenue (sum of market compensation and market penalty) versus aging with the performance evaluation implementation in Figure 4. Each dot is the mean value for the 10 random seeds. The scatter plot illustrates the tradeoff involved in adjusting the aging penalty w : higher aging penalties result in lower net revenues and lower aging. To avoid clutter, the dots in Figure 10 are labeled with the value of w , but the standard deviations are not shown. Figure 11 repeats the same plot without these labels and with shaded rectangles showing the standard deviation.

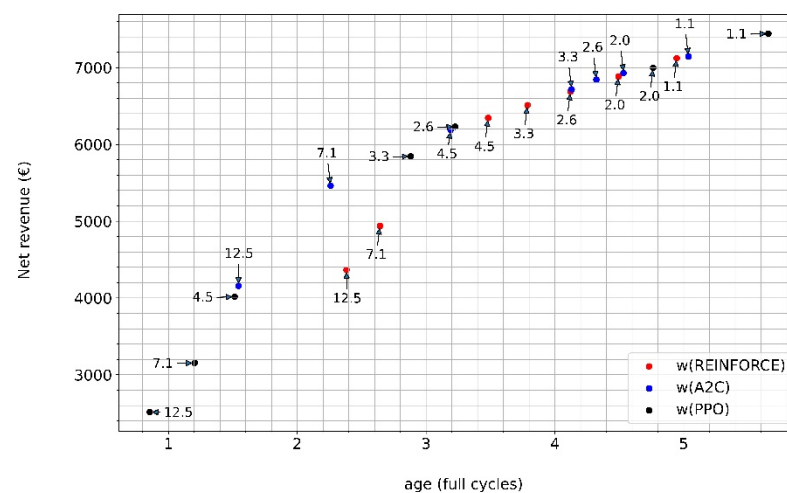


Figure 10. Scatter plot of net revenue and aging.

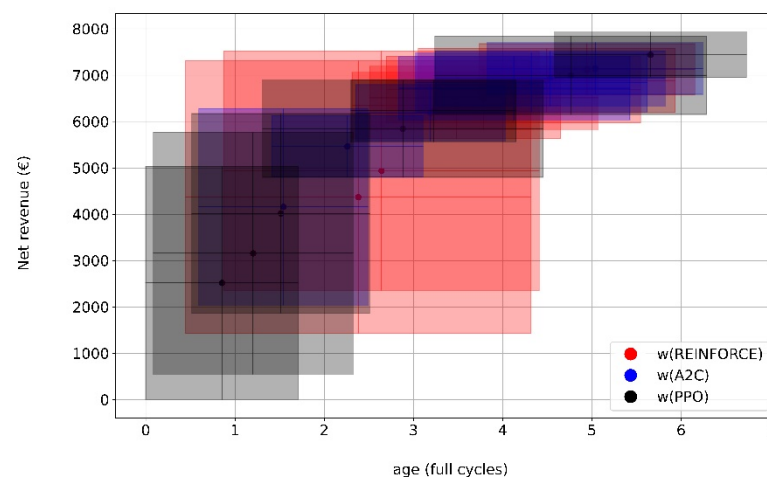


Figure 11. Scatter plot of net revenue and aging with standard deviation.

6. Discussion

The validation rewards in Figures 8 and 9 show that most of the learning occurs in the first 100 days. Only very minor improvements may be expected by continuing the training beyond the 1575 episodes used in this paper. The standard deviations in Figure 9 show that it is not possible to make any statistically significant statements about the superiority of any of the three RL algorithms.

Figures 10 and 11 illustrate the performance obtained by our agents with a realistic lithium-ion battery model. The dots are on a diagonal from the lower left to the upper right corner. This illustrates the tradeoff involved in adjusting the aging penalty w . A lower penalty results in higher net revenues and faster aging. This is to be expected intuitively since a lower value of w will decrease the negative aging penalty term in the reward function without affecting the positive market compensation term. As w is lowered, the positive market compensation term dominates the reward function and encourages the agent to take actions that increase the compensation. In other words, the agent is encouraged to bid higher capacities. According to Figure 4, higher capacities result in higher charging and discharging currents, which cause faster aging. Figure 10 shows that a straight line could be fitted to the dots with w values in the range 1.1–3.3. For higher w values, a significant drop in net revenues is observed. Our RL agent is not intended to be used in situations in which the aging penalty is very large compared to the market revenues. In such a situation, the business case for participating in the frequency reserve market is questionable.

The standard deviations in Figure 3 show that it is not possible to make statistically meaningful statements about the superiority of any of the RL algorithms. However, the diagonal trend discussed in the previous paragraph is also evident when the shaded areas are considered. Further, it is noted that when w is larger than 3.3, the shaded boxes are much larger. In the previous paragraph, 1.1–3.3 was identified as a relevant range. This can be due to the fact that with large w -values, the aging cost dominates the reward.

The relevant value for w depends on the actual cost of an equivalent full cycle of a particular battery. This cost depends greatly on raw material costs, supply chain disruptions, and government subsidies, which in turn can change drastically in response to global events such as pandemics and military conflicts. Thus, in this article, the aging penalty w is a parameter. If the aging cost value for a specific battery is known in terms of EUR per equivalent full cycle, the horizontal axis of Figure 10 can be converted to EUR by multiplying with this aging cost value. The w value can then be adjusted so that the difference between net revenue and aging cost is maximized.

The results are specific to the lithium-ion battery chemistry and the parameters of our case study battery presented in Tables 3 and 4. The methodology can be readily adapted to another lithium-ion battery by updating these parameters. The methodology can also be easily adapted to other battery chemistries by replacing the battery simulation block in Figures 3 and 4.

The results are specific to the Finnish FCR-N market. It is straightforward to generalize the approach to other auction-based frequency reserve markets in Finland or another country. The following modifications are needed. In the simulation environment, the current calculation should be according to the technical specification of the market, and the calculation of penalty minutes should be according to the market specification. The market price data and power grid frequency data used in our study have been obtained from the Finnish transmission system operator Fingrid's open data portal, so such data need to be obtained from the relevant TSO in another country. It is notable that the RL problem formulation does not need to be changed. Batteries can also be traded on other kinds of electricity markets. For example, a battery can perform energy arbitrage on day-ahead markets. It is not straightforward to generalize beyond frequency reserves markets to other auction-based electricity markets since significant changes to the RL problem formulation will be needed.

It is notable that RL practitioners generally use unique reward formulations, so it is not possible to make performance comparisons between different works. In this article,

a physics-based performance evaluation environment has been proposed which enables direct comparisons even with different reward formulations.

7. Conclusions

For each of the algorithms, learning was observed in the form of a reward that increased and eventually plateaued (Figures 6 and 8). The main statistical findings are summarized in Table 6. It is noted that for each of the algorithms, the reward is within the standard deviation of the other algorithms. It is concluded that each of the algorithms was successfully trained and that none of them was statistically superior to the others.

Table 6. Rewards and standard deviation at the last episode.

Algorithm	REINFORCE	A2C	PPO
Training reward	40.03	40.78	40.72
Training rewards std	15.17	17.74	14.05
Validation reward	1878.87	1917.16	1907.51
Validation reward std	68.03	101.43	187.03

As stated in Section 1, the contribution of this article is a methodology involving a realistic battery simulation to assess the performance of the trained RL agent with respect to battery aging in order to inform the selection of the weighting of the aging term in the RL reward formula. The results presented in Section 5.1 demonstrate that learning occurs and that none of the investigated RL algorithms is statistically superior to the others. As there is no statistically significant difference between the performance of the different algorithms, we conclude that the optimization problem was successfully addressed with all the algorithms. These kinds of results are frequently presented in the RL literature. Stopping the investigation at this point has two weaknesses. Firstly, it is not known if the RL agent trained in the RL environment could generalize to realistic battery dynamics if it would be tasked with managing a real battery. Secondly, since the results are quantified in terms of reward, they do not permit direct comparisons to other RL investigations of the same phenomenon, even if the battery parameters are identical if the RL reward formulations are different. Further, it is not possible to make performance comparisons to non-RL methods since the results are expressed in terms of the reward value.

To overcome these two weaknesses, a performance evaluation has been performed after confirming the learning on the validation dataset. The concept for the performance evaluation is presented in Figure 2, its implementation is presented in Figure 4, and the results are presented in Figures 10 and 11. The implementation in Figure 4 uses a realistic battery model and conforms to the technical specification of the Finnish FCR-N market. The results in Figures 10 and 11 are expressed in terms of net revenue and aging (equivalent full cycles). The battery parameters are presented in Tables 3 and 4 and the FCR-N market data and power grid frequency data are from the year 2020. Thus, any researcher is able to develop a battery FCR bidding optimizer, using either RL or non-RL methods, run them against this battery model and openly available market data and power grid data, and obtain results in terms of net revenue and aging that are directly comparable to the results that we have presented in Figures 10 and 11.

Author Contributions: Conceptualization, H.A. and S.S.; Data curation, H.A.; Formal analysis, H.A., S.S. and V.K.; Funding acquisition, S.S.; Investigation, H.A., M.P.-K. and V.V.; Methodology, H.A.; Project administration, S.S.; Resources, S.S.; Software, H.A.; Supervision, V.V.; Validation, H.A., S.S. and V.K.; Visualization, H.A.; Writing—original draft, S.S.; Writing—review & editing, H.A., S.S., V.K. and M.P.-K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Business Finland grant 7439/31/2018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Han, W.; Wik, T.; Kersten, A.; Dong, G.; Zou, C. Next-Generation Battery Management Systems: Dynamic Reconfiguration. *IEEE Ind. Electron. Mag.* **2020**, *14*, 20–31. [\[CrossRef\]](#)
2. Ben-Idris, M.; Brown, M.; Egan, M.; Huang, Z.; Mitra, J. Utility-Scale Shared Energy Storage: Business Models for Utility-Scale Shared Energy Storage Systems and Customer Participation. *IEEE Electr. Mag.* **2021**, *9*, 47–54. [\[CrossRef\]](#)
3. Zhang, Y.; Gevorgian, V.; Wang, C.; Lei, X.; Chou, E.; Yang, R.; Li, Q.; Jiang, L. Grid-Level Application of Electrical Energy Storage: Example Use Cases in the United States and China. *IEEE Power Energy Mag.* **2017**, *15*, 51–58. [\[CrossRef\]](#)
4. Tan, X.; Wu, Y.; Tsang, D.H.K. Pareto Optimal Operation of Distributed Battery Energy Storage Systems for Energy Arbitrage under Dynamic Pricing. *IEEE Trans. Parallel Distrib. Syst.* **2016**, *27*, 2103–2115. [\[CrossRef\]](#)
5. Attarha, A.; Amjady, N.; Dehghan, S. Affinely Adjustable Robust Bidding Strategy for a Solar Plant Paired with a Battery Storage. *IEEE Trans. Smart Grid* **2019**, *10*, 2629–2640. [\[CrossRef\]](#)
6. Filippa, A.; Hashemi, S.; Traholt, C. Economic Evaluation of Frequency Reserve Provision Using Battery Energy Storage. In Proceedings of the 2019 IEEE 2nd International Conference on Renewable Energy and Power Engineering (REPE), Toronto, ON, Canada, 2–4 November 2019; pp. 160–165.
7. Chen, H.; Baker, S.; Benner, S.; Berner, A.; Liu, J. PJM Integrates Energy Storage: Their Technologies and Wholesale Products. *IEEE Power Energy Mag.* **2017**, *15*, 59–67. [\[CrossRef\]](#)
8. Yuan, Z.; Zecchino, A.; Cherkaoui, R.; Paolone, M. Real-Time Control of Battery Energy Storage Systems to Provide Ancillary Services Considering Voltage-Dependent Capability of DC-AC Converters. *IEEE Trans. Smart Grid* **2021**, *12*, 4164–4175. [\[CrossRef\]](#)
9. Liu, Y.; Tang, Z.; Wu, L. On Secured Spinning Reserve Deployment of Energy-Limited Resources Against Contingencies. *IEEE Trans. Power Syst.* **2022**, *37*, 518–529. [\[CrossRef\]](#)
10. AlSkaif, T.; Crespo-Vazquez, J.L.; Sekuloski, M.; van Leeuwen, G.; Catalao, J.P.S. Blockchain-Based Fully Peer-to-Peer Energy Trading Strategies for Residential Energy Systems. *IEEE Trans. Ind. Inform.* **2022**, *18*, 231–241. [\[CrossRef\]](#)
11. Borowski, P.F. Zonal and Nodal Models of Energy Market in European Union. *Energies* **2020**, *13*, 4182. [\[CrossRef\]](#)
12. Mohsenian-Rad, H. Coordinated Price-Maker Operation of Large Energy Storage Units in Nodal Energy Markets. *IEEE Trans. Power Syst.* **2016**, *31*, 786–797. [\[CrossRef\]](#)
13. Vedullapalli, D.T.; Hadidi, R.; Schroeder, B. Combined HVAC and Battery Scheduling for Demand Response in a Building. *IEEE Trans. Ind. Appl.* **2019**, *55*, 7008–7014. [\[CrossRef\]](#)
14. Xu, B.; Zhao, J.; Zheng, T.; Litvinov, E.; Kirschen, D. Factoring the Cycle Aging Cost of Batteries Participating in Electricity Markets. In Proceedings of the 2018 IEEE Power & Energy Society General Meeting (PESGM), Portland, OR, USA, 5–10 August 2018; p. 1.
15. Akbari-Dibavar, A.; Mohammadi-Ivatloo, B.; Anvari-Moghaddam, A.; Nojavan, S.; Vahid-Ghavidel, M.; Shafie-khah, M.; Catalao, J.P.S. Optimal Battery Storage Arbitrage Considering Degradation Cost in Energy Markets. In Proceedings of the 2020 IEEE 29th International Symposium on Industrial Electronics (ISIE), Delft, The Netherlands, 17–19 June 2020; pp. 929–934.
16. Padmanabhan, N.; Ahmed, M.; Bhattacharya, K. Battery Energy Storage Systems in Energy and Reserve Markets. *IEEE Trans. Power Syst.* **2020**, *35*, 215–226. [\[CrossRef\]](#)
17. He, G.; Chen, Q.; Kang, C.; Pinson, P.; Xia, Q. Optimal Bidding Strategy of Battery Storage in Power Markets Considering Performance-Based Regulation and Battery Cycle Life. *IEEE Trans. Smart Grid* **2016**, *7*, 2359–2367. [\[CrossRef\]](#)
18. Guo, C.; Wang, X.; Zheng, Y.; Zhang, F. Optimal Energy Management of Multi-Microgrids Connected to Distribution System Based on Deep Reinforcement Learning. *Int. J. Electr. Power Energy Syst.* **2021**, *131*, 107048. [\[CrossRef\]](#)
19. Oh, E. Reinforcement-Learning-Based Virtual Energy Storage System Operation Strategy for Wind Power Forecast Uncertainty Management. *Appl. Sci.* **2020**, *10*, 6420. [\[CrossRef\]](#)
20. Natella, D.; Vasca, F. Battery State of Health Estimation via Reinforcement Learning. In Proceedings of the 2021 European Control Conference (ECC), Delft, The Netherlands, 29 June–2 July 2021; pp. 1657–1662.
21. Aaltonen, H.; Sierla, S.; Subramanya, R.; Vyatkin, V. A Simulation Environment for Training a Reinforcement Learning Agent Trading a Battery Storage. *Energies* **2021**, *14*, 5587. [\[CrossRef\]](#)
22. Huang, B.; Wang, J. Deep-Reinforcement-Learning-Based Capacity Scheduling for PV-Battery Storage System. *IEEE Trans. Smart Grid* **2021**, *12*, 2272–2283. [\[CrossRef\]](#)
23. Dong, Y.; Dong, Z.; Zhao, T.; Ding, Z. A Strategic Day-Ahead Bidding Strategy and Operation for Battery Energy Storage System by Reinforcement Learning. *Electr. Power Syst. Res.* **2021**, *196*, 107229. [\[CrossRef\]](#)
24. Knezevic, G.; Maligec, M.; Golub, V.; Topic, D. The Optimal Utilization of the Battery Storage for a Virtual Prosumer Participating on a Day-Ahead Market. In Proceedings of the 2020 International Conference on Smart Systems and Technologies (SST), Osijek, Croatia, 14–16 October 2020; pp. 155–160.
25. Liu, M.; Lee, W.-J.; Lee, L.K. Financial Opportunities by Implementing Renewable Sources and Storage Devices for Households Under ERCOT Demand Response Programs Design. *IEEE Trans. Ind. Appl.* **2014**, *50*, 2780–2787. [\[CrossRef\]](#)

26. Motta, S.; Aro, M.; Evens, C.; Hentunen, A.; Ikäheimo, J. A Cost-Benefit Analysis Of Large-Scale Battery Energy Storage Systems for Frequency Markets. In Proceedings of the CIRED 2021—The 26th International Conference and Exhibition on Electricity Distribution; Institution of Engineering and Technology, Online Conference, 20–23 September 2021; pp. 3130–3134.
27. Meng, Z.; Xu, Q.; Mao, T.; Li, M.; He, D. Research on the Full-Life-Cycle Operation Benefit Estimation of the Battery Energy Storage Station Anticipating the Ancillary Service Market in China. In Proceedings of the 2019 IEEE 3rd International Electrical and Energy Conference (CIEEC), Beijing, China, 7–9 September 2019; pp. 1086–1090.
28. Moolman, J.; Das, K.; Sørensen, P. Operation of Battery Storage in Hybrid Power Plant in Australian Electricity Market. In Proceedings of the 9th Renewable Power Generation Conference (RPG Dublin Online 2021), Institution of Engineering and Technology, Online Conference, 1–2 March 2021; pp. 349–354.
29. Goujard, G.; Badoual, M.D.; Janin, K.A.; Schwarz, S.; Moura, S.J. Optimal Siting, Sizing and Bid Scheduling of a Price-Maker Battery on a Nodal Wholesale Market. In Proceedings of the 2021 American Control Conference (ACC), Online Conference, 25–28 May 2021; pp. 602–607.
30. Lujano-Rojas, J.M.; Yusta, J.M.; Dominguez-Navarro, J.A.; Osorio, G.J.; Shafie-khah, M.; Wang, F.; Catalao, J.P.S. Combining Genetic and Gravitational Search Algorithms for the Optimal Management of Battery Energy Storage Systems in Real-Time Pricing Markets. In Proceedings of the 2020 IEEE Industry Applications Society Annual Meeting, Detroit, MI, USA, 10–16 October 2020; pp. 1–7.
31. Benini, M.; Canevese, S.; Cirio, D.; Gatti, A. Participation of Battery Energy Storage Systems in the Italian Balancing Market: Management Strategies and Economic Results. In Proceedings of the 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), Palermo, Italy, 12–15 June 2018; pp. 1–6.
32. Mohsenian-Rad, H. Optimal Bidding, Scheduling, and Deployment of Battery Systems in California Day-Ahead Energy Market. *IEEE Trans. Power Syst.* **2016**, *31*, 442–453. [[CrossRef](#)]
33. Robson, S.; Alharbi, A.M.; Gao, W.; Khodaei, A.; Alsaidan, I. Economic Viability Assessment of Repurposed EV Batteries Participating in Frequency Regulation and Energy Markets. In Proceedings of the 2021 IEEE Green Technologies Conference (GreenTech), Denver, CO, USA, 7–9 April 2021; pp. 424–429.
34. Khalilisenobari, R.; Wu, M. Impact of Battery Degradation on Market Participation of Utility-Scale Batteries: Case Studies. In Proceedings of the 2020 52nd North American Power Symposium (NAPS), Tempe, AZ, USA, 11–14 April 2021; pp. 1–6.
35. Gupta, A.; Vaishya, S.R.; Gupta, M.; Abhyankar, A.R. Participation of Battery Energy Storage Technologies in Co-Optimized Energy and Reserve Markets. In Proceedings of the 2020 21st National Power Systems Conference (NPSC), Gandhinagar, India, 17–19 December 2020; pp. 1–6.
36. Padmanabhan, N.; Bhattacharya, K. Including Demand Response and Battery Energy Storage Systems in Uniform Marginal Price Based Electricity Markets. In Proceedings of the 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA, 16–18 February 2021; pp. 1–5.
37. Kadri, A.; Mohammadi, F. Battery Storage System Optimization for Multiple Revenue Streams and Savings. In Proceedings of the 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), London, ON, Canada, 30 August–2 September 2020; pp. 1–5.
38. Prudhviraj, D.; Kiran, P.B.S.; Pindoriya, N.M. Stochastic Energy Management of Microgrid with Nodal Pricing. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 102–110. [[CrossRef](#)]
39. Chen, C.; Wei, Z.; Knoll, A.C. Charging Optimization for Li-Ion Battery in Electric Vehicles: A Review. *IEEE Trans. Transp. Electrif.* **2021**, *1*. [[CrossRef](#)]
40. Wu, J.; Wei, Z.; Li, W.; Wang, Y.; Li, Y.; Sauer, D.U. Battery Thermal- and Health-Constrained Energy Management for Hybrid Electric Bus Based on Soft Actor-Critic DRL Algorithm. *IEEE Trans. Ind. Inform.* **2021**, *17*, 3751–3761. [[CrossRef](#)]
41. Nefedov, E.; Sierla, S.; Vyatkin, V. Internet of Energy Approach for Sustainable Use of Electric Vehicles as Energy Storage of Prosumer Buildings. *Energies* **2018**, *11*, 2165. [[CrossRef](#)]
42. Ansari, M.; Al-Awami, A.T.; Sortomme, E.; Abido, M.A. Coordinated Bidding of Ancillary Services for Vehicle-to-Grid Using Fuzzy Optimization. *IEEE Trans. Smart Grid* **2015**, *6*, 261–270. [[CrossRef](#)]
43. Zeng, M.; Leng, S.; Maharjan, S.; Gjessing, S.; He, J. An Incentivized Auction-Based Group-Selling Approach for Demand Response Management in V2G Systems. *IEEE Trans. Ind. Inform.* **2015**, *11*, 1554–1563. [[CrossRef](#)]
44. Rassaei, F.; Soh, W.-S.; Chua, K.-C. Demand Response for Residential Electric Vehicles With Random Usage Patterns in Smart Grids. *IEEE Trans. Sustain. Energy* **2015**, *6*, 1367–1376. [[CrossRef](#)]
45. Zhou, M.; Wu, Z.; Wang, J.; Li, G. Forming Dispatchable Region of Electric Vehicle Aggregation in Microgrid Bidding. *IEEE Trans. Ind. Inform.* **2021**, *17*, 4755–4765. [[CrossRef](#)]
46. Tan, J.; Wang, L. A Game-Theoretic Framework for Vehicle-to-Grid Frequency Regulation Considering Smart Charging Mechanism. *IEEE Trans. Smart Grid* **2017**, *8*, 2358–2369. [[CrossRef](#)]
47. Donadee, J.; Ilic, M.D. Stochastic Optimization of Grid to Vehicle Frequency Regulation Capacity Bids. *IEEE Trans. Smart Grid* **2014**, *5*, 1061–1069. [[CrossRef](#)]
48. Sortomme, E.; El-Sharkawi, M.A. Optimal Charging Strategies for Unidirectional Vehicle-to-Grid. *IEEE Trans. Smart Grid* **2011**, *2*, 131–138. [[CrossRef](#)]

49. De Silva, D.; Sierla, S.; Alahakoon, D.; Osipov, E.; Yu, X.; Vyatkin, V. Toward Intelligent Industrial Informatics: A Review of Current Developments and Future Directions of Artificial Intelligence in Industrial Applications. *IEEE Ind. Electron. Mag.* **2020**, *14*, 57–72. [[CrossRef](#)]
50. Kahawala, S.; de Silva, D.; Sierla, S.; Alahakoon, D.; Nawaratne, R.; Osipov, E.; Jennings, A.; Vyatkin, V. Robust Multi-Step Predictor for Electricity Markets with Real-Time Pricing. *Energies* **2021**, *14*, 4378. [[CrossRef](#)]
51. Kempitiya, T.; Sierla, S.; de Silva, D.; Yli-Ojanperä, M.; Alahakoon, D.; Vyatkin, V. An Artificial Intelligence Framework for Bidding Optimization with Uncertainty in Multiple Frequency Reserve Markets. *Appl. Energy* **2020**, *280*, 115918. [[CrossRef](#)]
52. Beltrán, S.; Castro, A.; Irizar, I.; Naveran, G.; Yeregui, I. Framework for Collaborative Intelligence in Forecasting Day-Ahead Electricity Price. *Appl. Energy* **2022**, *306*, 118049. [[CrossRef](#)]
53. Haputhanthri, D.; de Silva, D.; Sierla, S.; Alahakoon, D.; Nawaratne, R.; Jennings, A.; Vyatkin, V. Solar Irradiance Nowcasting for Virtual Power Plants Using Multimodal Long Short-Term Memory Networks. *Front. Energy Res.* **2021**, *9*, 722212. [[CrossRef](#)]
54. Wood, D.A. Trend Decomposition Aids Short-Term Countrywide Wind Capacity Factor Forecasting with Machine and Deep Learning Methods. *Energy Convers. Manag.* **2022**, *253*, 115189. [[CrossRef](#)]
55. Pallonetto, F.; Jin, C.; Mangina, E. Forecast Electricity Demand in Commercial Building with Machine Learning Models to Enable Demand Response Programs. *Energy AI* **2022**, *7*, 100121. [[CrossRef](#)]
56. Cao, J.; Harrold, D.; Fan, Z.; Morstyn, T.; Healey, D.; Li, K. Deep Reinforcement Learning-Based Energy Storage Arbitrage With Accurate Lithium-Ion Battery Degradation Model. *IEEE Trans. Smart Grid* **2020**, *11*, 4513–4521. [[CrossRef](#)]
57. Wang, Y.; Lin, X.; Pedram, M. A Near-Optimal Model-Based Control Algorithm for Households Equipped With Residential Photovoltaic Power Generation and Energy Storage Systems. *IEEE Trans. Sustain. Energy* **2016**, *7*, 77–86. [[CrossRef](#)]
58. Chasparis, G.C.; Lettner, C. Reinforcement-Learning-Based Optimization for Day-Ahead Flexibility Extraction in Battery Pools. *IFAC-PapersOnLine* **2020**, *53*, 13351–13358. [[CrossRef](#)]
59. Lee, S.; Choi, D.-H. Energy Management of Smart Home with Home Appliances, Energy Storage System and Electric Vehicle: A Hierarchical Deep Reinforcement Learning Approach. *Sensors* **2020**, *20*, 2157. [[CrossRef](#)] [[PubMed](#)]
60. Lee, S.; Choi, D.-H. Reinforcement Learning-Based Energy Management of Smart Home with Rooftop Solar Photovoltaic System, Energy Storage System, and Home Appliances. *Sensors* **2019**, *19*, 3937. [[CrossRef](#)]
61. Yang, J.J.; Yang, M.; Wang, M.X.; Du, P.J.; Yu, Y.X. A Deep Reinforcement Learning Method for Managing Wind Farm Uncertainties through Energy Storage System Control and External Reserve Purchasing. *Int. J. Electr. Power Energy Syst.* **2020**, *119*, 105928. [[CrossRef](#)]
62. Liu, W.; Zhuang, P.; Liang, H.; Peng, J.; Huang, Z. Distributed Economic Dispatch in Microgrids Based on Cooperative Reinforcement Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 2192–2203. [[CrossRef](#)]
63. Zhou, S.; Hu, Z.; Gu, W.; Jiang, M.; Zhang, X.-P. Artificial Intelligence Based Smart Energy Community Management: A Reinforcement Learning Approach. *CSEE J. Power Energy Syst.* **2019**, *5*, 1–10. [[CrossRef](#)]
64. Wang, S.; Du, L.; Fan, X.; Huang, Q. Deep Reinforcement Scheduling of Energy Storage Systems for Real-Time Voltage Regulation in Unbalanced LV Networks With High PV Penetration. *IEEE Trans. Sustain. Energy* **2021**, *12*, 2342–2352. [[CrossRef](#)]
65. Qiu, X.; Nguyen, T.A.; Crow, M.L. Heterogeneous Energy Storage Optimization for Microgrids. *IEEE Trans. Smart Grid* **2016**, *7*, 1453–1461. [[CrossRef](#)]
66. Hofstetter, J.; Bauer, H.; Li, W.; Wachtmeister, G. Energy and Emission Management of Hybrid Electric Vehicles Using Reinforcement Learning. *IFAC-PapersOnLine* **2019**, *52*, 19–24. [[CrossRef](#)]
67. Zhou, J.; Xue, S.; Xue, Y.; Liao, Y.; Liu, J.; Zhao, W. A Novel Energy Management Strategy of Hybrid Electric Vehicle via an Improved TD3 Deep Reinforcement Learning. *Energy* **2021**, *224*, 120118. [[CrossRef](#)]
68. Hua, H.; Qin, Y.; Hao, C.; Cao, J. Optimal Energy Management Strategies for Energy Internet via Deep Reinforcement Learning Approach. *Appl. Energy* **2019**, *239*, 598–609. [[CrossRef](#)]
69. Muriithi, G.; Chowdhury, S. Optimal Energy Management of a Grid-Tied Solar PV-Battery Microgrid: A Reinforcement Learning Approach. *Energies* **2021**, *14*, 2700. [[CrossRef](#)]
70. Yang, T.; Zhao, L.; Li, W.; Zomaya, A.Y. Dynamic Energy Dispatch Strategy for Integrated Energy System Based on Improved Deep Reinforcement Learning. *Energy* **2021**, *235*, 121377. [[CrossRef](#)]
71. Chen, Z.; Hu, H.; Wu, Y.; Xiao, R.; Shen, J.; Liu, Y. Energy Management for a Power-Split Plug-In Hybrid Electric Vehicle Based on Reinforcement Learning. *Appl. Sci.* **2018**, *8*, 2494. [[CrossRef](#)]
72. Wu, J.; Wei, Z.; Liu, K.; Quan, Z.; Li, Y. Battery-Involved Energy Management for Hybrid Electric Bus Based on Expert-Assistance Deep Deterministic Policy Gradient Algorithm. *IEEE Trans. Veh. Technol.* **2020**, *69*, 12786–12796. [[CrossRef](#)]
73. Wei, Z.; Quan, Z.; Wu, J.; Li, Y.; Pou, J.; Zhong, H. Deep Deterministic Policy Gradient-DRL Enabled Multiphysics-Constrained Fast Charging of Lithium-Ion Battery. *IEEE Trans. Ind. Electron.* **2022**, *69*, 2588–2598. [[CrossRef](#)]
74. Karhula, N.; Sierla, S.; Vyatkin, V. Validating the Real-Time Performance of Distributed Energy Resources Participating on Primary Frequency Reserves. *Energies* **2021**, *14*, 6914. [[CrossRef](#)]
75. Giovanelli, C.; Kilkki, O.; Sierla, S.; Seilonen, I.; Vyatkin, V. Task Allocation Algorithm for Energy Resources Providing Frequency Containment Reserves. *IEEE Trans. Ind. Inform.* **2018**, *15*, 677–688. [[CrossRef](#)]
76. Giovanelli, C.; Sierla, S.; Ichise, R.; Vyatkin, V. Exploiting Artificial Neural Networks for the Prediction of Ancillary Energy Market Prices. *Energies* **2018**, *11*, 1906. [[CrossRef](#)]

-
77. Hua, H.; Qin, Z.; Dong, N.; Qin, Y.; Ye, M.; Wang, Z.; Chen, X.; Cao, J. Data-Driven Dynamical Control for Bottom-up Energy Internet System. *IEEE Trans. Sustain. Energy* **2022**, *13*, 315–327. [[CrossRef](#)]
 78. Chen, W.; Wu, N.; Huang, Y. Real-Time Optimal Dispatch of Microgrid Based on Deep Deterministic Policy Gradient Algorithm. In Proceedings of the 2021 International Conference on Big Data and Intelligent Decision Making (BDIDM), Guilin, China, 23–25 July 2021; pp. 24–28.
 79. Yi, Y.; Verbic, G.; Chapman, A.C. Optimal Energy Management Strategy for Smart Home with Electric Vehicle. In Proceedings of the 2021 IEEE Madrid PowerTech, Madrid, Spain, 28 June–2 July 2021; pp. 1–6.
 80. Opalic, S.M.; Goodwin, M.; Jiao, L.; Nielsen, H.K.; Lal Kolhe, M. A Deep Reinforcement Learning Scheme for Battery Energy Management. In Proceedings of the 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 23–26 September 2020; pp. 1–6.
 81. Sutton, R.; Barto, A. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2020.
 82. Omar, N.; Monem, M.A.; Firouz, Y.; Salminen, J.; Smekens, J.; Hegazy, O.; Gaulous, H.; Mulder, G.; van den Bossche, P.; Coosemans, T.; et al. Lithium Iron Phosphate Based Battery—Assessment of the Aging Parameters and Development of Cycle Life Model. *Appl. Energy* **2014**, *113*, 1575–1585. [[CrossRef](#)]
 83. Tremblay, O.; Dessaint, L.-A. Experimental Validation of a Battery Dynamic Model for EV Applications. *World Electr. Veh. J.* **2009**, *3*, 289–298. [[CrossRef](#)]
 84. MathWorks Battery. Available online: <https://www.mathworks.com/help/phymod/sps/powersys/ref/battery.html> (accessed on 16 May 2022).