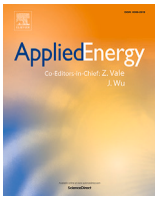
Applied Energy 388 (2025) 125561

Contents lists available at ScienceDirect

Applied Energy

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



# Deep reinforcement learning-based strategy for maximizing returns from renewable energy and energy storage systems in multi-electricity markets

Javier Cardo-Miota[a,b], Hector Beltran[a], Emilio Pérez[a], Shafi Khadem[b,*], Mohamed Bahloul[b,c]

[a] Department of Industrial Systems Engineering and Design, Universitat Jaume I, Castelló de la Plana, Spain
[b] International Energy Research Center, Tyndall National Institute, UCC, Cork, T12 RC5P Ireland
[c] Water & Energy Transition Unit, Vlaamse Instelling voor Technologisch Onderzoek (VITO), Mol, B-2400 Belgium

## HIGHLIGHTS

- Deep reinforcement learning algorithms for bidding strategies.
- Decision-making algorithm for participating in multiple electricity markets.
- Operation of renewable energy systems co-located with battery energy storage systems.
- Introduces a versatile framework that can be applied universally.

## ARTICLE INFO

*Keywords:*
Deep reinforcement learning
Markov decision process
Bidding strategy
Battery energy storage system management
Renewable energy systems
Multi-electricity markets participation

## ABSTRACT

The integration of Renewable Energy Sources (RES) with Energy Storage Systems (ESS) presents challenges and opportunities in optimizing their participation in electricity markets. This study introduces a novel approach that leverages Deep Reinforcement Learning (RL) algorithms to develop optimal bidding strategies for collocated RES with Battery ESS (BESS) configurations, enabling multi-market participation in both energy and ancillary services (AS) markets. The proposed method uses a Markov Decision Process (MDP) framework to manage BESS utilization dynamically, considering market conditions and technical constraints. As an RL agent, the actor–critic approach known as the Twin Delayed Deep Deterministic (TD3) Policy Gradient algorithm is implemented. A data-driven training process facilitates model learning while minimizing the required training dataset and time. Focused on the Irish context, the case study involves participation in both the day-ahead energy market and reserve services for frequency droop curve response of the DS3 Programme. Historical data from a 7 MW solar PV plant and a 1 MWh BESS are utilized to evaluate the performance. The RL agent dynamically adapts to market dynamics and system constraints, achieving substantial economic benefits compared to benchmark strategies, with an additional 8271€, 166,738€, and 11,369€, respectively.

## 1. Introduction

In recent years, there has been a significant increase in the adoption of renewable energy systems (RESs) globally. To exemplify, note that the RES capacity additions reached around 473 GW in 2023, representing an increase of nearly 50 % compared to 2022 [1]. This growth is expected to continue, driven by ambitious goals set by governments worldwide to achieve Net Zero Emissions by 2050 (NZE Scenario) [2]. However, while this significant expansion in RES generation promises environmental benefits, it also poses various challenges for power systems and Transmission System Operators (TSOs) [3].

The variability of renewable generation, influenced by natural factors, introduces uncertainty and fluctuations into the system [4]. This can lead to grid stability problems, resulting in more and more deviations between generation and demand, causing over-voltages, frequency events, and even blackouts etc. [5–7]. In this context, the Ancillary Services (ASs) [8] emerge as crucial solutions. Complementary to elec-

* Corresponding author.
*Email addresses:* jcardo@uji.es (J. Cardo-Miota), hbeltran@uji.es (H. Beltran), pereze@uji.es (E. Pérez), shafi.khadem@tyndall.ie (S. Khadem), mohamed.bahloul@vito.be (M. Bahloul).

https://doi.org/10.1016/j.apenergy.2025.125561
Received 20 August 2024; Received in revised form 30 December 2024; Accepted 14 February 2025
Available online 9 March 2025

tricity markets, these services immediately take on the responsibility of mitigating these imbalances to ensure a secure and high-quality energy supply for consumers. Their main function is to constantly balance generation and demand, compensating for unforeseen variations and maintaining the frequency and voltage of the system within acceptable limits in real-time. The ASs are classified into four main categories: frequency response, capacity, voltage control and black-start capability [8]. Traditionally, only conventional generators have been entitled to provide ASs due to their flexible and dispatchable generation over time. However, the gradual decline in their presence in electricity markets, coupled with the significant environmental impact associated with their activities, has prompted the emergence of new, more sustainable and flexible providers [9]. Examples include Active Demand Management (ADM) and Energy Storage Systems (ESSs), both of which offer the necessary flexibility to meet the requirements of such services [10].

Among them, the Battery Energy Storage Systems (BESSs) are crucial solutions due to their technical capabilities, such as rapid response times, efficient energy supply and absorption, and long-lasting operational performance [11]. Combining BESSs with RESs, including wind farms and photovoltaic (PV) power plants, has become a widely adopted approach to address the instability caused by RES generation while enhancing the overall system flexibility. The main objectives of integrating these systems are to reduce generation intermittency resulting from unpredictable weather patterns, streamline system dispatch operations, minimize energy losses, and improve overall system efficiency. Moreover, this integration aims to enhance system flexibility and participation across different energy market segments. The significant decrease in BESS prices further supports this approach, as observed in recent years [12].

However, despite these advantages, installing BESSs still requires a substantial initial investment, posing challenges to their economic feasibility. Numerous studies have concluded that participating solely in energy markets by doing arbitrage (charging at low prices to discharge at high prices) is not profitable for BESSs [13,14]. This underscores the need to optimize their participation across various energy and AS electricity markets to ensure their economic viability.

Numerous studies regarding the multi-market participation of BESS, both standalone or working in parallel with RES power plants, can be found in the literature. In [15], the authors propose a novel approach based on linear matrix inequalities with a semi-definite programming model to define the optimal scheduling decisions of a Li-ion BESS that provides peak shaving, and power factor improvement services while minimizing the electricity bill of a consumer house in the Australian power system. The results clearly outperform the decisions made with a simple linear model, reducing the energy losses, and improving the charging and discharging efficiencies. The authors of [16] introduce two chance-constrained optimization methods for both the day-ahead and intraday scheduling of a PV power plant with a BESS. This system actively participates in the day-ahead (DA) and intraday (ID) energy markets while also providing Primary Frequency Regulation (PFR) service in the Italian power system. The findings demonstrate a robust optimization approach capable of effectively managing uncertainties derived from forecast errors. The authors in [17] propose a multi-stage stochastic programming model to define the optimal bidding strategy of a wind farm combined with a BESS that simultaneously participates in the DA, ID energy markets and in the secondary frequency reserve (SFR) control market within the Iberian electricity market. Through a comparative economic analysis, they demonstrate the effectiveness of their optimization algorithm, achieving a substantial increase in profitability when the BESS is allowed to participate in the SFR control market. Furthermore, some of the authors of this work recently proposed a Mixed-Integer Linear Programming (MILP) algorithm in [18] to optimize the scheduling of a combined PV–BESS power plant. The aim was to minimize PV power plant clipping losses while providing grid ancillary services within the Irish Delivering a Secure, Sustainable Electricity System (DS3) Programme. The results unequivocally demonstrate that offering DS3 services can serve as the main source of profit for such systems.

In short, there are numerous additional studies based on classical optimization algorithms related to BESS operation, participating in several markets [19–21]. However, the current landscape of electricity markets is characterized by a highly dynamic and fluctuating environment, both in terms of prices and generation, primarily due to the presence of RESs, as well as on the demand side. This volatility is further increased by the intricate interaction of various market layers. In such a complex environment, traditional linear and quadratic optimization methods, which rely on model-based algorithms such as MILP, may be less appropriate to address this problem.

In recent years, Reinforcement Learning (RL), a sub-field of Machine Learning (ML), primarily associated with control problems in robotics and video games, has emerged as an alternative to classical decision-making algorithms thanks to its adaptability to changing environments and easy implementation based on rewards [22]. This ML branch, based on Artificial Neural Networks (ANN) and a Markov Decision Process (MDP), is characterized by achieving optimal decision-making through a trial-and-error training process among agent–environment interactions [22]. The effectiveness of RL in BESS operation, both standalone and combined with PV power plants, has been extensively demonstrated in various research articles. The authors of [23] introduce a novel RL-based model for defining bidding strategies of four BESS, participating in both an energy and an Automatic Generation Control (AGC) market using a Function Approximation based on RL (FARL) algorithm. The outcomes show robust training and higher revenue from AGC market than traditional Q-learning and State-Action-Reward-State-Action (SARSA) methods. [24] addresses a bidding optimization for a standalone BESS in the Australian energy and Frequency Control Ancillary Services (FCAS) markets using a Proximal Policy Optimization (PPO) RL model. The results of this model show notably higher profitability than strategies that focus solely on one electricity market. In [25], four RL models are compared to optimize the capacity scheduling of a PV plant with BESS in the PJM market, with PPO delivering the highest economic performance. [26] proposes a methodology for bidding in real-time energy markets, using a generalized MDP to address RES–BESS facility constraints, achieving superior revenue generation compared to arbitrage-focused strategies.

Although previous research exists, to the best of the authors' knowledge, there is a noticeable gap in analyzing an operational strategy for a collocated RES–BESS installation participating simultaneously in at least three day-ahead electricity markets, including both energy and AS markets. Table 1 highlights the specific research gaps that this work addresses in comparison to prior studies. This study takes a significant step forward by presenting a novel RL framework for optimizing the operation of RES–ESS in these multi-market environments. Unlike prior studies, the adopted methodology here is adaptable to the unique market structures of different regions, which are often shaped by local and international regulations. The novel contribution of this research is threefold: First, it introduces a versatile framework that can be applied universally, regardless of the electric market context or RES installation specifics. Second, to the best of the authors' knowledge, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm has been introduced for the first time within this domain, utilizing dual critic neural networks to refine return estimations. Third, we propose an innovative non-linear battery degradation model that dynamically links the degradation of the battery to the RL-agent's decisions, adding a more sophisticated and realistic layer to the optimization process. This model enables accurate tracking of degradation based on the operational choices of the agent, allowing it to learn strategies that effectively reduce long-term degradation. This enhancement not only captures real-world degradation patterns more accurately but also underscores the advantage of

**Table 1**
Research gaps addressed in this work.

| Reference | RES–BESS installation | Day-ahead electricity markets | Adaptability to different RES | Non-linear degradation model | Adaptability to different regions | 3 electricity markets | TD3 agent |
|---|---|---|---|---|---|---|---|
| [23] | ✓ | ✓ | | ✓ | | | |
| [24] | ✓ | ✓ | | | | | |
| [25] | ✓ | ✓ | | | | | |
| [26] | ✓ | | ✓ | | | | |
| This work | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

an RL-based approach, as traditional linear or MILP optimization methods are less equipped to handle such complex, non-linear degradation dynamics.

The primary goal of RL optimizer is to maximize profitability while adhering to operational constraints, offering a practical and adaptable approach to understanding and managing RES–BESS systems in fluctuating electricity markets. Also, a robust sequential Markov Decision Process (MDP) environment has been developed to simulate the dynamics of RES–BESS systems in various market conditions accurately. Finally, building on previous research [18,27], this work demonstrates the efficacy of the model using data from the Irish electricity market and a PV power plant in Ireland. The TD3 agent, a central innovation of this work, is adept at navigating the complexities of modern power systems and operates within a continuous action space to enhance decision-making efficiency.

To sum up, the key contributions and the novelty of this paper are:

- MDP Framework: An MDP framework is introduced to optimize strategic policies for maximizing profits in the collocated RES–BESS plant.
- Actor–Critic Algorithm: An Actor–Critic policy gradient algorithm is proposed that operates in a continuous action space, improving control precision and adaptability to environmental changes.
- LSTM Neural Network: A Long Short-Term Memory (LSTM)-based neural network is implemented to discern complex temporal patterns within the action space, enhancing the system's adaptability to dynamic environmental conditions.
- Non-Linear Degradation Model: A novel non-linear battery degradation model is incorporated, dynamically linking battery degradation to operational decisions based on the depth of discharge evolution. This model introduces a realistic layer to the optimization, capturing the degradation impact of specific cycles and depths that traditional linear models fail to address. This aspect significantly improves the model's ability to represent real-world degradation, allowing the RL agent to learn strategies that minimize battery wear and extend BESS lifespan.
- Training Process Control: The training process is refined by employing a data-driven approach to expedite learning and reduce training duration.
- Versatile Sequential Environment: The applied versatility of this environment facilitates its application in various RES technologies and market structures, thereby increasing its applicability and robustness.

The rest of the paper is organized as follows: Section 2 provides a comprehensive description of the system under consideration, including the electricity market context, detailed RES and BESS models, and operational constraints. Section 3 presents the problem formulation, introducing the MDP framework developed and the TD3 agent algorithm. Results and discussions are presented in Section 4, where the case study and the simulation setup are also explained. Finally, concluding remarks are offered in Section 5.

## 2. System description

The considered RES power plant combined with a BESS is illustrated in Fig. 1. A DC tightly coupled configuration is adopted here as it meets the technical and economic requirements of this type of installation [28].

Note that the system's voltages are automatically regulated by the energy conversion systems (DC/DC and DC/AC) to optimize the production of the power plant, as well as by the transformer, as depicted in Fig. 1. Apart from that, the constraints related to the power grid's currents are implicitly incorporated into the formulation by defining the nominal operating power of the DC/DC and the DC/AC conversion systems. Recall that the analysis of the impact of power injection from the plant on the connection point (such as power quality and short-circuit analysis) lies beyond the scope of our analysis.

The market options considered for the RES–BESS hybrid plant and the operating model of each system element are described below.

### 2.1. Electricity market context

For a multi-electricity market participation problem, it is crucial to clearly define the market options for each element of the installation. In this context, we outline the market framework, which is explained in the following. Note that the proposed strategy is designed for participation in day-ahead markets (both energy and AS markets), where

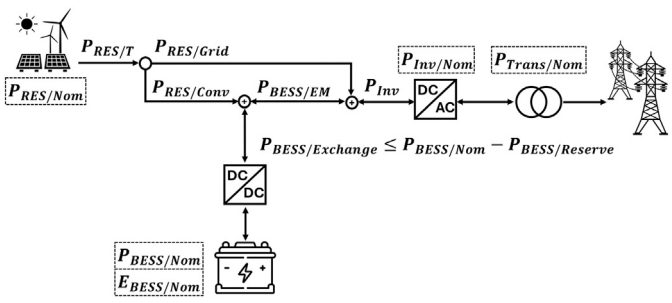



**Fig. 1.** RES–BESS plant electric scheme.

the bidding strategies for both RES and BESS must be defined the day before.

Firstly, we assume that part of the energy generated by the RES power plant is sold to what we define as the RES energy market. The benefits derived from this energy sale at each time-step are represented by Equation (1), where $Price_{RES}(k)$ is the price of the RES energy market at time period $k$ and $P_{RES/Grid}(k)$ is the power generated by the RES power plant that is directly discharged to the grid in this time step. The RES energy market could be a wholesale energy market, encompassing the day-ahead energy market, an auction within a discrete intraday energy market, or a round within a continuous intraday energy market. However, participating directly in the wholesale energy market may not be as advantageous for a RES power plant as initially expected. The increasing penetration of solar PV plants within electrical power systems is reshaping the market landscape: prices often drop during daylight hours, only to rise sharply during peak demand periods when these technologies disconnect. This price erosion during daylight hours can potentially disrupt the business model of PV plants. Consequently, many choose to participate in the market through a bilateral contract called Power Purchase Agreements (PPAs), which ensures a fixed selling price for the energy produced. This shields them from market volatility and helps ensure profitability.

$$B_{RES}(k) = Price_{RES}(k) \cdot P_{RES/Grid}(k) \cdot \Delta k \tag{1}$$

Secondly, it is necessary to determine the market participation of the BESS. In this sense, participating in energy markets is essential to take advantage of the price difference between off-peak and peak hours. Therefore, it is decided to incorporate the BESS into the wholesale energy market, either in the day-ahead or intraday energy market. This enables charging from both the RES generation and the grid, with subsequent discharge directly into the grid. Equation (2) defines the benefits or costs obtained by the BESS participating in the energy market at each time step. Here, $Price_{EM}(k)$ represents the price of the energy market at period $k$, and $P_{BESS/EM}(k)$ represents the power exchanged by the BESS in the energy market, either purchasing energy ($P_{BESS/EM}(k) < 0$) or selling it ($P_{BESS/EM}(k) > 0$). Note that whether it represents a cost or a profit depends on the sign of the variable $P_{BESS/EM}(k)$.

$$B_{BESS/EM}(k) = Price_{EM}(k) \cdot (P_{BESS/EM}(k)) \cdot \Delta k \tag{2}$$

Finally, we propose to complement the arbitrage strategy by involving the BESS in an AS market to enhance profitability. The flexible operation of BESSs, coupled with their technical characteristics, enables them to respond to frequency events over an extended period. These frequency services are often remunerated based on power availability, generating profits by reserving power without significant wear and tear, making it a highly appealing market for this type of technology. Thus, the participation of the BESS in a reserve AS market is assumed. Equation (3) indicates the benefits obtained by the BESS through its participation in this AS market in each time step, where $Price_{AS}(k)$ is the price of the AS market at period $k$ and $P_{BESS/Reserve}(k)$ is the power reserved to participate in this market in this time period.

$$B_{BESS/AS}(k) = Price_{AS}(k) \cdot P_{BESS/Reserve}(k) \cdot \Delta k \tag{3}$$

Note that, in this work, we made the assumption that the prices $Price_{RES}(k)$, $Price_{EM}(k)$, and $Price_{AS}(k)$ for each time period $k$ are known at the time of running the optimization.

### 2.2. Renewable energy system model

In this study, the versatility of the proposed framework is extended to encompass any RES installation. This adaptability allows the optimization tool to easily adapt to the variability in generation from different renewable sources. Whether wind farms exploit wind speed or solar PV plants capture the sunlight, the framework remains unconcerned with respect to the energy source, ensuring its applicability to several renewable energy technologies combined with a BESS. This flexibility not only expands the scope of application but also underscores the robustness of the tool in managing the dynamic nature of renewable energy generation.

Among the multiple options, the RES installation model considered in this study is formally expressed by Equation (4). Accordingly, the power generated by the RES plant ($P_{RES/T}(k)$) can either be sold to the grid ($P_{RES/Grid}(k)$) or used to charge the battery ($P_{RES/Conv}(k)$) through the DC–DC converter.

$$P_{RES/T}(k) = P_{RES/Grid}(k) + P_{RES/Conv}(k) \tag{4}$$

### 2.3. Battery energy storage system model

When considering the BESS, two different power components are identified: the first corresponds to the actual power exchanged by the battery (represented by $P_{BESS/Exchange}(k)$) and the second represents the power reserved to provide the AS (represented by $P_{BESS/reserve(k)}$).

Regarding $P_{BESS/Exchange}(k)$, this can be: negative, indicating that the battery is charging; positive, implying that the battery is discharging; or zero, indicating that the battery is idle. The charging process of the battery can occur either by purchasing energy from the energy market ($P_{BESS/EM}(k) < 0$) or through the RES generation ($P_{RES/Conv}(k)$), while the discharge can only occur by selling it in the energy market ($P_{BESS/EM}(k) > 0$). Thus, the energy stored in the battery ($E_{BESS}(k)$) is subject to operational constraints defined by Equation (5), where $\eta$ represents the battery efficiency. The value of $\eta$ varies depending on whether the battery is charging ($\eta = \eta_{charge}$) or discharging ($\eta = \eta_{discharge}$).

$$E_{BESS}(k+1) = E_{BESS}(k) + P_{BESS/Exchange}(k) \cdot \eta \cdot \Delta k \tag{5}$$

Regarding the second power component, $P_{BESS/reserve(k)}$, note that together with $E_{BESS/Reserve(k)}$ stand for the power and capacity reserved for participation in the AS. This implies that, for each hourly period $k$, both power and capacity of the BESS for discharge must be reserved to address a frequency droop.

Additionally, to ensure the committed participation of BESS in both markets, $E_{BESS}(k)$ and $E_{BESS/Reserve}(k)$, must fall within limits as indicated in Equations (6) and (7), where $E_{Min}$ and $E_{Max}$ represent the lower and upper bounds.

$$E_{Min} \leq E_{BESS}(k) - E_{BESS/Reserve}(k) \leq E_{Max} \tag{6}$$

$$E_{Min} \leq E_{BESS}(k) + E_{BESS/Reserve}(k) \leq E_{Max} \tag{7}$$

Alternatively, the constraints on the stored energy of the battery can be expressed as a function of the State of Charge ($SOC$), which defines the percentage of available storage capacity of the battery relative to its maximum capacity ($SOC(k) = E_{BESS}(k)/E_{max}$). Thus, Equations (5), (6), and (7) are rewritten as equations (8), (10) respectively:

$$SOC(k+1) = SOC(k) + \frac{P_{BESS/Exchange}(k) \cdot \eta \cdot \Delta k}{E_{max}} \tag{8}$$

$$SOC_{Min} \leq SOC_{BESS}(k) - SOC_{BESS/Reserve}(k) \leq SOC_{Max} \tag{9}$$

$$SOC_{Min} \leq SOC_{BESS}(k) + SOC_{BESS/Reserve}(k) \leq SOC_{Max} \tag{10}$$

### 2.4. Grid and operational system constraints

Several additional constraints must also be taken into account. The battery charging, defined by $P_{Res/Conv}(k)$ and $P_{BESS/EM}(k)$, is restricted by two factors: the nominal power of the battery ($P_{BESS/Nom}$), the available charging capacity of the battery ($SOC_{Av/Charge}(k) = SOC_{Max} -$

$SOC(k)$), and the nominal power of the converter ($P_{Conv/Nom}$). To define these technical limitations, a virtual variable named $P_{Max}(k)$ is introduced, representing the maximum power among the defined powers. Thus, Equation (11) defines the technical constraint related to charging, where $P_{Max/Charge}(k)$ is defined in Equation (12). Note that Equation (12) is our way to address Equation (10).

$$|P_{BESS/Exchange}(k)| \leq P_{Max/Charge}(k), \text{ when } P_{BESS/Exchange}(k) < 0 \quad (11)$$

$$P_{Max/Charge}(k) = \min(P_{Conv/Nom}, P_{BESS/Nom}, \frac{SOC_{Av/Charge}(k) \cdot E_{max}}{\Delta k}) \quad (12)$$

Similarly, the discharge of the battery, along with the power reserved for providing the AS, is constrained by the nominal power of the battery ($P_{BESS/Nom}$), by the available discharge capacity of the battery ($SOC_{Av/Discharge}(k) = SOC(k) - SOC_{Min}$), and the nominal power of the converter ($P_{Conv/Nom}$). Equations (13) and (14) define these discharge technical constraints. Analogously to Equation (12), Equation (14) is our way to address Equation (9).

$$P_{BESS/Exchange}(k) + P_{BESS/Reserve}(k) \leq P_{Max/Discharge}(k), \text{ when } P_{BESS/Exchange}(k) > 0 \quad (13)$$

$$P_{Max/Discharge}(k) = \min(P_{Conv/Nom}, P_{BESS/Nom}, \frac{SOC_{Av/Discharge}(k) \cdot E_{max}}{\Delta k}) \quad (14)$$

Furthermore, the total power exchanged with the grid, together with the power reserved for the AS provision, is limited by the inverter, which has the same nominal power as the transformer ($P_{Inv/Nom} = P_{Trans/Nom}$), as shown in Equation (15):

$$P_{RES/Grid}(k) + P_{BESS/EM}(k) + P_{BESS/Reserve}(k) \leq P_{Inv/Nom} \quad (15)$$

### 2.5. BESS degradation model

Battery degradation plays a crucial role in the overall viability of energy storage systems. As such, integrating a battery degradation model into the control and decision-making framework is of utmost importance. Recall that the battery lifetime is influenced by two primary factors: cycling ageing and calendar ageing. The cycling ageing is related to BESS operational conditions, such as SOC, DOD, etc. In contrast, calendar ageing is associated mainly with non-operational conditions, and its impact on control performance is often less significant than that of cycling ageing. Consequently, recent research suggests omitting calendar ageing effects in control framework design to avoid additional complexity in the decision algorithms [29,30].

In this paper, a nonlinear degradation model is proposed to integrate the control framework. For simplicity, the ageing model is focused on the cycling influence and derived from recent work [18]. However, in order to enhance the model's reliability, a DOD cycling cost is introduced to better model the nonlinear relationship between the number of cycles and their DOD.

The model is also based on a cost per cycle index. The cycling cost ($C_{cyc}$) is measured in €/kWh and calculated using the equation (16), where $C_{BESS}$ represents the total cost of the BESS in €, $E_{max}$ is the installed BESS capacity in kWh, $DOD_{max}$ denotes the maximum DOD recommended by the manufacturer, and $N_{cycles}$ is the estimated number of cycles achievable at $DOD_{max}$.

$$C_{cyc} = \frac{C_{BESS}}{E_{max} \cdot DOD_{max} \cdot N_{cycles}} \quad (16)$$

To take into account a cycling cost dependent on the operation decided by the RL agent for the BESS, a dynamic cycling cost is considered.

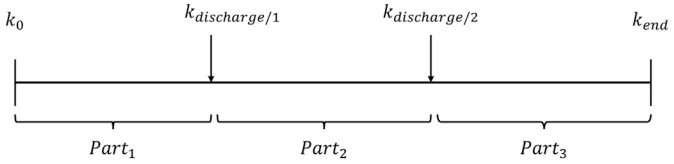



**Fig. 2.** Parts of the day according to the allowed discharges.

This is introduced by equation (17), and calculated by multiplying the base cycling cost $C_{cyc}$ by a factor $z_{factor}(k)$, which varies depending on the value of an also dynamic (*DOD(k)*).

$$C^F_{cyc}(k) = z_{factor}(k) \cdot C_{cyc} \quad (17)$$

This *DOD(k)* is calculated as the difference between the maximum and minimum SOC experienced by the BESS during different parts of the day, as shown in equation (18). To leverage price spreads in the energy market, only two arbitrage opportunities are allowed per day: one in the period with the highest energy market price in the morning ($k_{discharge/1}$), and another in the evening period with highest price ($k_{discharge/2}$). Outside of these peak price times, the battery is only allowed to charge. In this way, the day is divided into three different parts based on permissible discharge events, as illustrated in Fig. 2.

$$DOD(k) = SOC_{max}(part_i) - SOC_{min}(part_i) \quad (18)$$

Within each part of the day, the agent dynamically identifies $SOC_{max}(part_i)$ and $SOC_{min}(part_i)$ at each time step ($k$) based on its operational decisions. This process enables the calculation of *DOD(k)* at each time step, reflecting the influence of the agent's decisions on the cycling cost.

Then, the value of $z_{factor}(k)$ is determined by the depth of discharge registered at each time step within each part of the day. To modelize the influence of an increasing DOD on a higher degradation, we categorize the $z_{factor}(k)$ into six intervals. Such intervals try to encapsulate the different degradation rates observed in the cycle life curve for varying the DOD percentages (Fig. 3). As illustrated, the battery life decreases non-linearly when the DOD increases, following a pattern where higher DOD levels correspond to fewer achievable cycles. This relationship is captured through the following set of values for $z_{factor}(k)$, as shown in equation (19).

$$z_{factor}(k) = \begin{cases} 1 & \text{if } DOD(k) > 70\% \\ 0.75 & \text{if } DOD(k) \in (55\% - 70\%) \\ 0.50 & \text{if } DOD(k) \in (40\% - 55\%) \\ 0.40 & \text{if } DOD(k) \in (30\% - 40\%) \\ 0.10 & \text{if } DOD(k) \in (10\% - 30\%) \\ 0.02 & \text{if } DOD(k) < 10\% \end{cases} \quad (19)$$

This categorization was established based on the cycle life degradation curve provided by SAFT Batteries for lithium iron phosphate (LFP) cells at a reference temperature of 25∘C [31]. From the curve, we observe that as DOD decreases, the cycle life increases significantly. For example:

- At $DOD(k) \in (70\% - 80\%)$, approximately, 5000 cycles are achievable, which we use a baseline for $z_{factor}(k) = 1$.
- At $DOD(k) = 40\%$, the battery can complete around 10,000 cycles, justifying a reduction factor to reflect the double cycle life in this range ($z_{factor} = 0.5$).
- At $DOD(k) = 10\%$, the cycle life reaches up to 200,000 cycles, indicating a much lower degradation rate, which we approximate with a $z_{factor} = 0.02$ to reflect this forty-fold increase compared to the baseline operational range.

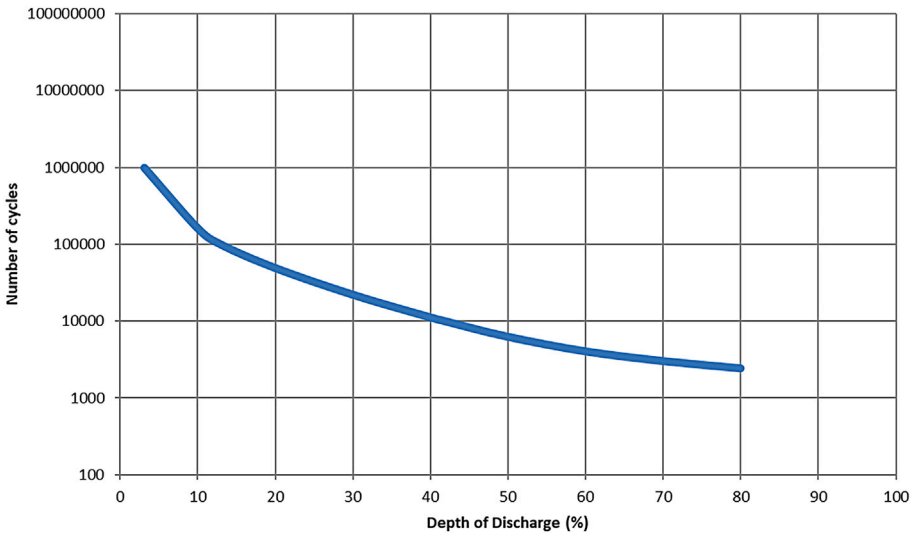



**Fig. 3.** Curve showing the relationship between the number of cycles and *DOD* [31].

Thus, higher values of *DOD(k)* lead to larger $z_{factor}(k)$ values, increasing the cycling cost in the RL-based decision-making model. This approach effectively captures the non-linear impact of the DOD on the battery degradation, aligning the cycling cost with realistic degradation patterns observed in LFP cells.

Overall, the total degradation penalty assumed in this RL-based model is calculated as indicated in equation (20).

$$Degradation_{penalty}(k) = P_{BESS/Exchange}(k) \cdot \frac{C^{F}_{cyc}(k)}{2} \tag{20}$$

## 3. Methodology

### 3.1. RL background

RL represents a cutting-edge domain within the field of ML, offering a powerful framework for addressing dynamic optimization and control problems. RL is based on a simple but effective decision-making paradigm, in which an agent interacts with an environment, learning to make optimal decisions through a process of trial and error. Unlike traditional supervised learning methods where labeled datasets are used to determine specific actions, RL allows autonomous decision-making that rewards desirable behaviors and penalizes suboptimal ones based on the feedback received from the environment. This feedback loop promotes continuous learning and adaptation, allowing RL agents to navigate complex, uncertain environments and derive optimal strategies for various tasks.

The MDP provides a formal framework for modeling RL problems. As shown in Fig. 4, within this framework, the agent interacts continuously with the environment by selecting actions, and the environment responds by transitioning to new states. Each interaction yields a numerical reward, representing the outcome of the agent's decisions. The main objective of this optimization strategy is to maximize the expected return, which is the cumulative sum of rewards obtained by the agent over time until the last observation.

At each time step $k$, the agent obtains the observation $s(k)$, also known as the state, from the environment. Using this information, the agent's policy, denoted by $\pi$, selects the corresponding actions $a(k) = \pi(s(k))$. In the next time step, the agent receives a numerical reward $r(k+1)$ and the subsequent observation $s(k+1)$ as a consequence of the impact of the action on the environment. This interaction $(s(k), a(k) \rightarrow r(k+1), s(k+1))$ is known as a trajectory and is repeated N times until the final state $s_N$. This sequence of steps, from the initial state to the final state, is called an episode.


**Fig. 4.** Agent–environment interactions in Reinforcement learning problems.

The main goal of the agent is to maximize the expected return from each state $s(k)$ until the end of the episode $s_N$. This expected return is also known as the value function $V_\pi(s(k))$, which estimates the total reward the agent can expect to receive from that state onward, considering its current policy. Equation (21) defines the value function, where $\gamma \in (0, 1)$ is the discount factor that reflects the uncertainty in future rewards. Alternatively, some approaches aim to maximize the Q-value function, as defined in Equation (22). The Q-value represents the expected return specifically for a state-action pair, providing insight into the potential cumulative reward of taking a particular action in a given state.

$$V_\pi(s) = E[\sum_{i=0}^{N} \gamma^i \cdot r(k+i+1) | s(k) = s] \tag{21}$$

$$Q_\pi(s,a) = E[\sum_{i=0}^{N} \gamma^i \cdot r(k+i+1) | s(k) = s, a(k) = a] \tag{22}$$

To optimize this value or Q-value function, the goal of the agent is to determine the optimal policy ($\pi^*$) for selecting the best action in each state. This requires a balance during the learning process between exploration, where the agent takes random actions to explore the environment and gather information, and exploitation, where it selects actions that maximize its expected return.

### 3.2. Markov decision process formulation

Fig. 5 illustrates the information flow diagram connecting the different elements in an MDP model. This is formed by:

#### 3.2.1. Environment

The environment comprehends all the elements accepted as external to the agent that influences the desired reward to be maximized.

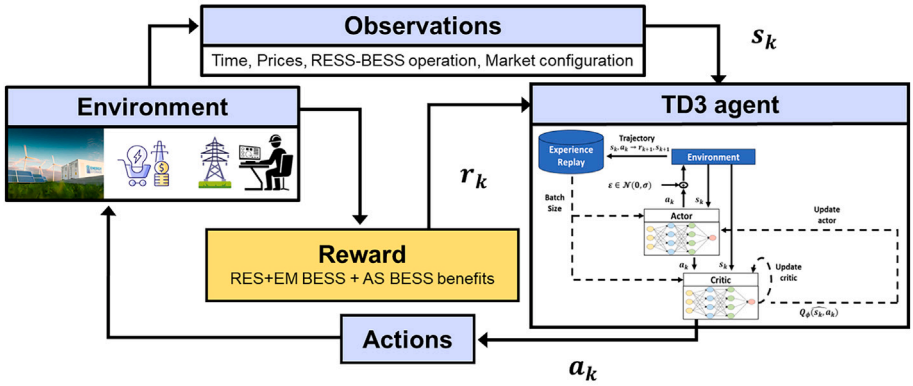



**Fig. 5.** Proposed Markov Decision Process for the multi-market participation of a RES–BESS installation.

Modeling an appropriate environment is crucial to facilitate the learning process of the RL model.

In this study, the environment encompasses all the market information along with the operation of the RES–BESS plant. The environment is modeled based on the definition of the control variables of the battery operation: $P_{BESS/Exchange}(k)$ and $P_{BESS/Reserve}(k)$. Additionally, by adopting a sequential environment [32], it is possible to effectively address the different system constraints. Thus, the proposed control sequence involves first defining $P_{BESS/Exchange}(k)$ and then $P_{BESS/Reserve}(k)$.

To manage the information effectively, we implemented a slotted time model where decisions and system constraints are defined for discrete time intervals, denoted by $k$, of fixed and equal length. Consequently, each day is divided into 24 time slots, each with a duration of 1 hour.

Within each time slot, the price signals from the RES energy market ($Price_{RES}(k)$), the energy market ($Price_{EM}(k)$), and the reserve AS market ($Price_{AS}(k)$) are collected.

Additionally, as explained before, only two arbitrage opportunities per day are allowed to take full advantage of the price spread in the energy market: one in the period with the maximum energy market price during the morning time ($k_{discharge/1}$), and another during the evening period with maximum energy market price ($k_{discharge/2}$). In contrast, the battery is allowed to charge during the rest of the periods, giving priority to charging from RES sources whenever available.

For each time period $k$, it is determined whether it is either a discharging ($k = k_{discharge/1}$ or $k = k_{discharge/2}$) or a charging period ($k \neq k_{discharge/1}$ and $k \neq k_{discharge/2}$).

During discharge periods, the variable $P_{Max}(k)$ is calculated according to Equation (14). Then, $P_{BESS/Exchange}(k)$ is defined by Equation (23), where $\delta = [0, 1]$ defines the percentage of $P_{Max}(k)$ allocated for $P_{BESS/Exchange}(k)$:

$$P_{BESS/Exchange}(k) = +\delta(k) \cdot P_{Max/Discharge}(k) \tag{23}$$

Furthermore, since all the battery discharge energy is sold in the energy market, it is derived that $P_{BESS/EM}(k) = P_{BESS/Exchange}(k)$. Additionally, during discharge periods, all the RES generation is fed into the grid $P_{RES/Grid}(k) = P_{RES/T}(k)$, and $P_{RES/Conv}(k) = 0$.

During charge periods, $P_{Max/Charge}(k)$ is defined according to Equation (12). Then, $P_{BESS/Exchange}(k)$ is determined by Equation (24).

$$P_{BESS/Exchange}(k) = -\delta(k) \cdot P_{Max/Charge}(k) \tag{24}$$

To determine the source of the charging power, priority is given to charging from RES. The following cases apply:

- When $P_{RES/T}(k) \geq |P_{BESS/Exchange}(k)|$, all charging power comes from RES ($P_{RES/Conv}(k) = |P_{BESS/Exchange}(k)|$ and $P_{BESS/EM}(k) = 0$), and the surplus RES generation is fed into the grid ($P_{RES/Grid}(k) = P_{RES/T}(k) - |P_{BESS/Exchange}(k)|$).
- When $P_{RES/T}(k) < |P_{BESS/Exchange}(k)|$, all RES generation is used to charge the battery ($P_{RES/Conv}(k) = P_{RES/T}(k)$ and $P_{RES/Grid}(k) = 0$), and the remainder is charged from the grid ($P_{BESS/EM}(k) = -(|P_{BESS/Exchange}(k)| - P_{RES/T}(k))$).

Once the power exchanged by the battery and the management of the power generated by the RES plant are defined, the SOC of the BESS is updated using Equation (8).

Next, the available power of the battery ($P_{BESS/Av}(k)$), and that of the converter ($P_{Conv/Av}(t)$) are calculated to define the power reserved for the AS market. Again, the following cases apply:

- If the battery is charging, the available power of the battery, and converter are $P_{BESS/Av}(t) = P_{BESS/Nom}$, and $P_{Conv/Av}(t) = P_{Conv/Nom}$, respectively.
- If the battery is discharging, the available powers are $P_{BESS/Av}(t) = (P_{BESS/Nom} - P_{BESS/Exchange}(t))$, and $P_{Conv/Av}(t) = (P_{Conv/Nom} - P_{BESS/Exchange}(t))$, respectively.

Subsequently, participation in the reserve AS market is defined. To do this, the new maximum discharge power is recalculated according to Equation (25):

$$P'_{Max/Discharge}(k) = \min(P_{Conv/Av(t)}, P_{BESS/Av(k)}, , \frac{(SOC(k+1) - SOC_{Min}) \cdot E_{max}}{\Delta k}) \tag{25}$$

Then, $P_{BESS/Reserve}(k)$ is defined by Equation (26), where $\beta = [0, 1]$ defines the percentage of $P_{Max}(k)$ allocated for it:

$$P_{BESS/Reserve}(k) = +\beta(k) \cdot P'_{Max/Discharge}(k) \cdot \eta_{Discharge} \tag{26}$$

Considering the power limitation associated with the nominal capacity of the inverter is crucial here. In order to save costs, we defined an inverter with a rated power capacity 25 % lower than the peak output power of the RES plant ($P_{Inv/Nom} = 0.75 \cdot (P_{RES/Nom} + P_{BESS/Nom})$) [33]. Thus, ensuring that the RES power generation fed into the grid must comply with Equation (15) is important.

### 3.2.2. Observation space

The observation space, also known as state, gathers all the relevant information from the environment for the agent to take informed actions. In this context, the 17 environment observations listed in Equation (27) are identified to be introduced to the agent at each time step. These include:

- Three states involving the multiple seasonalities of the operation (the *period of the day*, *day of the week*, and *season of the year*).

- Six regarding price and RES–BESS operation (the different electricity prices, the RES generation data, the *SOC(k)*, and the *DOD(k)*).
- Eight describing the market situation on a specific day. These include the maximum price of the energy market during the morning ($Price_{EM/\max/1}$) and afternoon ($Price_{EM/\max/2}$) and their corresponding periods ($k_{EM/maxPrice/1}$, $k_{EM/maxPrice/2}$, respectively); the minimum price of the energy market ($Price_{EM/\min}$) and its corresponding period ($k_{EM/minPrice}$); and the minimum and maximum price of the AS market ($Price_{AS/\min}$, $Price_{AS/\max}$, respectively).

$$\begin{aligned} S = \{ & s(k) = (k, weekday, season, Price_{EM}(k), Price_{AS}(k), \\ & Price_{RES}(k), P_{RES/T}(k), SOC(k), DOD(k), \\ & Price_{EM/\max/1}, Price_{EM/\max/2}, k_{EM/maxPrice/1}, k_{EM/maxPrice/2}, \\ & Price_{EM/\min}, k_{EM/minPrice}, Price_{AS/\min}, Price_{AS/\max})\} \end{aligned} \quad (27)$$

### 3.2.3. Action space

The actions involve the control variables over which the agent makes decisions to maximize the reward. Thus, depending on the current state *s(k)*, the agent selects the action *a(k)*.

We identify two control variables: $P_{BESS/Exchange}(k)$ and $P_{BESS/Reserve}(k)$. In this regard, as modeled in the environment, the control elements for these variables are the utilization percentages of the maximum power defined as $\delta$ and $\beta$, respectively.

However, to facilitate the learning of the RL model, we opt to simplify the decision-making for the agent by reducing the number of decisions to take. We assume that the BESS will be fully used every period, either for charging/discharging, reserve services, or both concurrently. To ensure this, $\beta$ is set to 1. Consequently, by adjusting the parameter $\delta$, it is possible to determine the battery power allocation both for exchanging and reserving.

Due to the RL approach used, a continuous action space is selected. This implies that the number of possible action values the agent can take within the action range is infinite. Thus, the action space is defined as in Equation (28).

$$A = \{a(k) = \delta(k) | \delta(k) \in [0, 1]\} \quad (28)$$

### 3.2.4. Reward function

The reward function returns a numerical result that explains the consequences of the agent–reward interaction. The main goal of the optimization strategy is to maximize the expected return, understood as the sum of the rewards obtained by the agent over a given period of time. Once the action *a(k)* at the state *s(k)* is taken, the agent completes the trajectory by interacting with the environment to obtain the next state *s(k+1)* and the reward *r(k+1)*.

In this study, the reward function, defined in Equation (29), reflects the financial benefits of the RES–BESS system at each time step. These benefits arise from three sources: benefits from the RES system ($B_{RES}(k)$), benefits from BESS participation in the energy market ($B_{BESS/EM}(k)$), benefits from BESS participation in the reserve AS market ($B_{BESS/AS}(k)$), and the degradation penalty calculated in Section 2.5 ($Degradation_{penalty}(k)$).

$$r_k = B_{RES}(k) + B_{BESS/EM}(k) + B_{BESS/AS}(k) - Degradation_{penalty}(k) \quad (29)$$

## 3.3. Twin Delayed Deep Deterministic Policy Gradient (TD3) agent

The agent is central to RL models as it is the one taking the actions to maximize rewards. In this study, we employed an agent based on the TD3 algorithm, which is an advanced variant of the Deterministic Policy Gradient (DDPG) method in RL [34]. It stands out for its emphasis on reducing the Q-value overestimation and improving the training stability.

The TD3 algorithm belongs to the Actor–Critic family in RL and employs two neural networks, as illustrated in Fig. 5. The actor-network serves as the policy function ($\pi_\theta(s, a)$), selecting an action for every state. It takes the state *s(k)* as input and outputs the action *a(k)*. The TD3 agent assumes a deterministic policy, meaning the policy always selects the action that maximizes the return at each state. The actor requires the Q-value function estimated by the critic network to select the optimal action, ensuring that the chosen action maximizes the expected return. To encourage exploration, a normally distributed noise signal is added to the actor's output, enabling the selection of actions that deviate from those maximizing the expected return. Furthermore, the TD3 agent uses a continuous action space, allowing for finer control and adaptability, providing precise adjustments to dynamic changes in the environment and a more accurate representation of continuous decision variables.

In contrast, the critic network evaluates this policy using the Q-value function estimation ($\widehat{Q}_\phi(s, a)$). The critic network takes both the state *s(k)* and action *a(k)* as inputs, producing the expected discounted return of the trajectory *Q(s,a)* as output. The accurate estimation of the Q-value by the critic is crucial for the correct implementation of the control action, as the entire optimization process relies on this estimated return.

As illustrated in Fig. 5, at each time step $k$ during training, the resulting trajectory ($s(k), a(k) \rightarrow r(k+1), s(k+1)$) is stored in an *Experience Replay Buffer*. Once enough trajectories are accumulated in the buffer, a random minibatch of $T$ trajectories is selected to update both the actor and critic networks. Since the networks operate independently, their parameters $\theta$ and $\phi$ are optimized separately.

The actor-network, aiming to select the action that maximizes the expected return in each state, updates its parameters using the Stochastic Gradient Ascent (SGA) rule to adjust the policy's output towards the direction of maximum Q-value growth. The learning rate $\alpha$ determines the magnitude of parameter updates, influencing the speed and stability of the learning process, as illustrated in Equation (30).

$$\theta(k+1) = \theta(k) + \alpha \nabla Q(s(k), \pi_\theta(s(k))) \quad (30)$$

In contrast, the loss function for the critic network is defined in Equation (31).

$$\frac{1}{|T|} \sum_{i=1}^{T} (r(k+1) + \gamma \widehat{Q}(s(k+1), a(k+1)|\phi) - \widehat{Q}(s(k), a(k)|\phi)^2 \quad (31)$$

Unlike deep learning techniques, where the actual value is compared with the prediction to measure improvement in the loss function, the critic in RL compares two Q-value estimations because the expected return is not known in advance. This technique is known as *bootstrapping*. If both estimations are calculated by the same neural network updated simultaneously, no improvement can be observed every time the loss function is computed. To solve this problem and stabilize the learning process, the concept of target neural networks is introduced [35]. These target neural networks are copies of the main Q-network but are updated less frequently, providing more stable target values ($r(k+1) + \gamma \widehat{Q}(s(k+1), a(k+1)|\phi')$) for the learning update rule.

The TD3 agent introduces a unique feature regarding this target neural network of the critic: it employs two target critic networks (twin critics). These assess the action values more accurately by mitigating the overestimation of the Q-value function in certain states with several possible actions. The TD3 agent considers the more conservative Q-value estimate ($r(k+1) + \gamma \min_{1,2}(\widehat{Q}(s(k+1), a(k+1)|\phi))$) [36].

Other improvements of the TD3 agent include:

- **Delayed policy updates**: deterministic policy gradient methods heavily depend on accurate Q-value estimations for action selection. During early training iterations, poor Q-value estimations can destabilize policy training. To address this, the policy network is updated less frequently than the critic networks, reducing the variance of policy updates and leading to more stable learning.
- **Target policy smoothing**: this technique addresses overfitting to narrow peaks in the Q-function due to poor critic estimations, promoting smoother and more stable policy updates. Some normal noise

is added to the action predicted by the target policy to generate a set of perturbed actions. The target Q-value is then calculated using the minimum Q-value of these perturbed actions.

Furthermore, we implement an LSTM layer into the actor and critic networks to handle time series observations. LSTM networks allow capturing complex temporal patterns in the input data, which is crucial for dynamic environments such as power systems [37]. This capability allows the RL algorithm to adapt to fluctuations in RES generation and energy market dynamics while retaining relevant long-term information for informed decision-making by the RL agent. By incorporating LSTM layers into the actor and critic networks, the RL algorithm improves its ability to model and predict power system behavior, leading to more efficient and cost-effective decisions in RES–BESS operation.

Finally, to provide further clarity, the steps involved in the TD3 algorithm's process for solving the RES–BESS optimization problem can be summarized as follows:

1. **State Observation**: At each time step $k$, the agent observes the current state of the environment, which includes information such as the current power generation, market prices, and energy storage status.
2. **Action Selection**: Based on the observed state, the actor network selects an action ($P_{BESS/Exchange}(k)$) that aims to maximize the expected future reward. A noise term is added to encourage exploration.
3. **Reward Calculation**: After executing the action, the agent receives a reward based on the profit generated by the RES–BESS power plant, considering electricity market dynamics and system constraints.
4. **Critic Network Update**: The critic networks evaluate the action by estimating the Q-value ($\hat{Q}_{\phi}(s,a)$). The TD3 algorithm mitigates overestimation by using two critic networks and selecting the minimum Q-value.
5. **Actor Network Update**: The actor network is updated using the Q-value provided by the critic network, adjusting its policy to maximize future rewards.
6. **Experience Replay Buffer**: The agent stores the transition $s(k), a(k) \rightarrow r(k+1), s(k+1))$ in a replay buffer. Periodically, a batch of experiences is sampled to update both the actor and critic networks, improving the stability of the learning process.
7. **Target Networks**: To further stabilize the training, target networks for both the actor and critics are employed. These networks are updated less frequently, preventing the learning process from becoming unstable due to noisy updates.
8. **Delayed Policy Updates**: The policy (actor) is updated at a slower rate than the critic to ensure more stable learning during early training iterations, where the Q-value estimates can be inaccurate.

These steps are repeated throughout the training process until the agent converges to an optimal policy that efficiently manages the operation of the RES–BESS system across multiple electricity markets.

## 4. Case study, simulation and results

### 4.1. Case study

The case study is used to evaluate the efficacy of the proposed RL-based scheduling algorithm when applied to an RES power plant integrating BESS.

The technical features of the power plant under consideration are as follows:

- A solar PV power plant is considered as the RES installation. Historical PV generation hourly data obtained from a real 7 MW rated PV plant is used.
- A BESS with the following specifications is integrated: a capacity of 1 MWh, nominal power of 1 MW, charging efficiency ($\eta_{charge}$) of 95 %, discharging efficiency ($\eta_{discharge}$) of 95 %, minimum SOC of 10 %, and maximum SOC of 90 %.
- The nominal power of both the DC–DC converter and the AC–DC inverter is 1 MW and 6 MW, respectively.
- The nominal power of the transformer is 6 MW.

To calculate the cycling cost $C_{cyc}$, we assume a BESS cost of 200 €/kWh as indicated in [38], with a maximum DOD of 80 % and a cycle life of 5000 cycles at this DOD level [18]. Based on these parameters, we obtain a cycling cost of $C_{cyc} = 50$€/MWh.

To participate in the electricity market, although the results can be applied to other regions while adjusting the markets' parameters, we decided to connect the PV–BESS power plant to the Irish power system. Thus, the electricity markets assumed are:

- **RES energy market participation**: a PPA contract offering a rate of is 73€/MWh [18].
- **BESS energy market**: the Irish Day-Ahead market (DAM) is considered the trading platform where the BESS engages in energy arbitrage. This is because it represents the largest share of energy trading in Ireland, reaching 85.83 % in 2023 [39].
- **BESS reserve AS market**: Within the Irish context, the AS markets are included in the Delivering a Secure, Sustainable Electricity System (DS3) Programme [40]. Although BESSs are allowed to participate in this programme, their scope is limited to providing services related to the frequency droop curve response within a 20-minute time frame [41]. These services include Fast Frequency Response (FFR) and the following reserve services: the Primary Operating Reserve (POR), Secondary Operating Reserve (SOR), Tertiary 1 Operating Reserve (TOR 1), and Tertiary 2 Operating Reserve (TOR 2) [40].

Note that given that the DS3 programme functions in half-hourly intervals, each time step in our analysis requires considering two DS3-related periods.

These services are procured through two types of procedures: Volume Capped (VC) and Volume Uncapped (VU) [40]. In this study, we assume participation in the DS3 using the VU procurement method. The VU procurement process is characterized by the absence of limits on the total volume of services procured, with regulated tariffs being applied. It is important to highlight that within VU procurement, the activation of the FFR, the reserve, and the ramping services occur exclusively as a response to events of frequency decline rather than during periods of over-frequency [40].

The trading period payment for the participation in the services under the VU procedure is calculated as in Equation (32):

$$\text{Trading Period Payment} = \text{Available Volume} \cdot \text{Payment Rate}) \cdot \cdot \text{Scaling Factor} \cdot \text{Trading Period Duration} \quad (32)$$

where the available volume is the amount of power (MW) based on the absolute lowest sustainable value that the unit is capable of achieving in the given timeframe for the service. The Payment Rate terms for each service are defined in the DS3 System Service Statement of Payments and included in Table 2. The Scaling Factor is a term dependent on several scalars related to the response of service providers themselves and to external conditions affecting the power system. In the case of a BESS that participates in the 5 services related to the frequency droop curve response, the scaling factor will be composed of the multiplication of the following scalars [42]:

1. **Temporal Scarcity Factor**: Reflects the situation of the system in terms of System Non-Synchronous Penetration (SNSP) ratio. It is classified into different clusters based on the SNSP level, and their corresponding values are shown in Table 2. In this work, we assume perfect knowledge of the SNSP ratio, and consequently, of the Temporal Scarcity Factor.

**Table 2**
Payment Rates and Temporal Scarcity Scalar for the DS3 system services depending on the SNSP interval [43,44].

| System Service | Pay Unit | Payment Rate (€) | Temporal Scarcity Scalar | | | |
|---|---|---|---|---|---|---|
| | | | *SNSP* < 50 % | 50 % ≤ *SNSP* < 60 % | 60 % ≤ *SNSP* < 70 % | 70 % ≤ *SNSP* |
| POR | MWh | 2.92 | 1 | 1 | 4.7 | 6.3 |
| SOR | MWh | 1.76 | 1 | 1 | 4.7 | 6.3 |
| TOR 1 | MWh | 1.40 | 1 | 1 | 4.7 | 6.3 |
| TOR 2 | MWh | 1.12 | 1 | 1 | 4.7 | 6.3 |
| FFR | MWh | 1.94 | 0 | 1 | 0 | 6.3 |

2. **Performance Scalar**: It depends on the result of the performance assessment conducted by the TSOs. Its value is between 0 and 1. A unit value of this scalar is assumed in this work due to the lack of detailed information.
3. **Product Scalar**: It depends on the reserve trigger type and reserve trigger scalar compared to 49.3 Hz. Its value ranges between 0 and 1. A unit value of this scalar is assumed in this work due to the lack of detailed information.
4. **Locational Scalar**: The minimum value equals 1 and reflects the geographical location of the providing unit. A unit value of this scalar is assumed in this work due to the lack of detailed information.
5. **FFR Fast Response Scalar**: Scales how fast the provided response is. Due to the rapid response provided by batteries, we assume that the BESS can achieve an FFR response time of less than 0.15 s, resulting in an FFR Fast Response scalar of 3.
6. **FFR Continuous Provision Scalar**: It is set to 1.5 if the unit is available to support the system for POR, SOR, TOR1, and FFR during the trading period, otherwise it is set to 1. We have decided to participate continuously in the five frequency response services (FFR, POR, SOR, TOR 1, and TOR 2) with the same committed power. Therefore, we assume an FFR Continuous provision scalar equal to 1.5.

#### 4.1.1. Simulation setup

For the experiments, 2022 data from the DA market and the SNSP ratio extracted from [45,46] were used. In addition, the PV generation data is actual data provided directly by the PV plant owner. To mitigate the considerable computational cost and time involved in the training process, the model was trained on a limited dataset consisting of only 20 days. This dataset was carefully sampled to include 5 days from each season of the year. Similarly, for testing purposes, another set of 20 days was employed, again sampling 5 days from each season. Furthermore, to provide a comprehensive evaluation of the optimization, a separate test using data from the entire year, excluding the training dataset, was conducted to assess its effectiveness.

Furthermore, the TD3 agent is configured as follows. The actor-network architecture consists of an input layer with 17 features corresponding to the state space dimension. Note that the data is normalized using the min–max normalization technique, scaling it to a range between -1 and 1 by subtracting the minimum value and dividing by the range of the data. This transformation helps the models better understand the relationships between the different inputs. The input layer is followed by an LSTM layer with 64 units and a ReLU activation function. Subsequently, two hidden layers with 64 neurons each, also employing ReLU activation functions, are included before the output layer. The output layer consists of a hidden layer with dimensions matching the actor's action space (1), using a tanh activation function.

The critic network takes two inputs: the observation space (17 features) and the action space (1 feature). After concatenating both inputs, the network architecture is similar to that of the actor-network featuring: an LSTM layer with 64 units and a ReLU activation function, followed by two hidden layers with 64 neurons each, also using ReLU activation functions. Subsequently, the output layer consists of a single neuron in a hidden layer, which predicts the expected Q-value until the end of the episode.

**Table 3**
RES–BESS multi-market participation case study: Agent hyperparameters.

| Hyperparameter | Value | Hyperparameter | Value |
|---|---|---|---|
| Batch Size | 128 | Exploration standard deviation | 0.5 |
| Experience buffer length | $10^5$ | Target standard deviation | 0.5 |
| Discount Factor | 1 | Target update frequency | 2 |
| Actor learning rate | $5 \cdot 10^{-}5$ | Policy update frequency | 2 |
| Critic learning rate | $5 \cdot 10^{-}5$ | Polyak averaging factor | 0.01 |

In terms of the RL general parameters, the number of episodes was set to 20, aligning with the number of training days, each comprising 24 samples to optimize the cumulative reward within a day. These episodes were repeated for 100 epochs. A Bayesian optimization approach was employed to determine the optimal agent hyperparameters, with the selection criterion being the maximization of the Average Reward attained at the end of the training process. The resulting hyperparameters are outlined in Table 3.

The proposed algorithms and simulations are implemented with *Matlab*® using a laptop with an Intel Core i7 processor and an NVIDIA GeForce RTX 3060 GPU. The total training time required was 4 hours and 36 minutes.

Fig. 6 shows the training process of the TD3 agent using the dataset from the first 5 days. Note that the blue line represents the return obtained in each episode, while the dashed black line indicates the moving average return over 25 episodes, providing a clearer trend of the agent's learning progression. Initially, the agent explores the environment by selecting random actions to achieve different rewards. Consequently, the returns obtained during these early episodes tend to be lower, and in some cases even negative (note that the reward function has been standardized to improve the actor and critic neural networks' understanding). Starting around episode 180, the agent has learned the dynamics of the environment and selects actions that maximize the return for each episode. As a result, the return values consistently exceed 15, leading to a significant increase in the average return compared to the earlier episodes.

#### 4.1.2. Results and discussion

Note that two scenarios are proposed as benchmarks to compare and evaluate the performance of the algorithm. In the first one, we assume that the entire capacity of the BESS is reserved exclusively for providing DS3 services (PV + DS3 approach). In this case, as the battery does not undergo any charge–discharge cycling for energy arbitrage ($P_{BESS/Exchange} = 0$), no degradation penalty is considered in this approach. In the second scenario, the BESS is used solely for arbitrage in the DA market (PV + DA approach). In the latter case, the BESS is fully discharged during peak price periods, both in the morning and afternoon, while being charged during the off-peak hour of each day and from PV generation during the daylight hours between discharges. Furthermore, a DDPG agent is also trained for comparison purposes. This agent shares the same neural network architectures and hyperparameters as the TD3 agent, and it is trained using the same dataset.

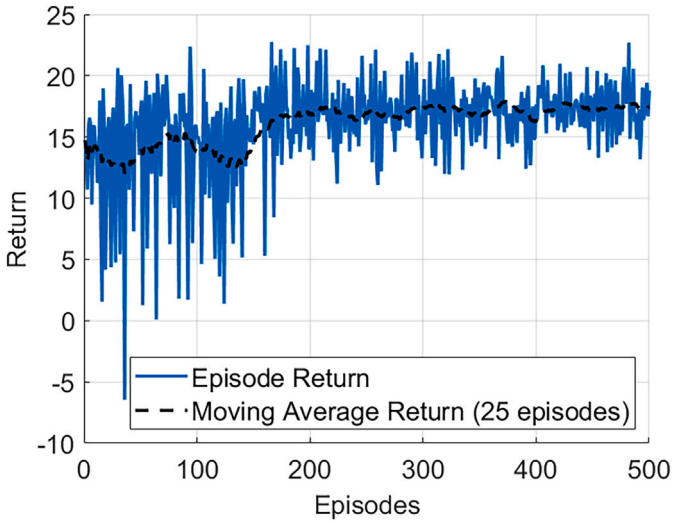



**Fig. 6.** Training process of the TD3 agent.

Table 4 shows the profits obtained by the different approaches using the 20 days of testing. The results confirm the robustness of the proposed RL algorithm and the effectiveness of the training process. In all scenarios, the multi-service strategy implemented with the TD3 optimization consistently outperforms the other approaches, including the DDPG agent proposal. While the DDPG agent improves upon the PV + DA and PV + DS3 benchmarks in winter and summer, it underperforms in spring and autumn, likely due to higher variability in PV generation and market prices during these periods. In contrast, the TD3 agent shows consistent performance throughout all seasons, demonstrating the capability of the agent to make sound decisions regardless of the period of the year. Particularly noteworthy are the summer days, when the TD3 agent yields additional benefits of 974€ and 1100€ over the PV + DA and PV + DS3 approaches, respectively, and 630€ more than the DDPG agent. In addition, note that our proposal consistently achieves lower degradation costs than the PV + DA case, highlighting that the TD3 agent has learned to mitigate the battery degradation during the training process. Furthermore, observe that our model also outperforms the profits obtained in the PV + DS3 scenario, even though this approach does not account for any economic penalties related to the battery degradation. Overall, the application of the proposed TD3 agent during the 20 days of testing results in total additional revenues of 5982€ and 1573€ for the PV + DA and PV + DS3 approaches, respectively, and 1749€ more than the DDPG agent.

The performance of the RL-based controller is then analyzed over 5 consecutive days through Figs 7, 8, and 9. Note that, although the model operates on an hourly time step basis, the figures are displayed in 30-minute intervals to accommodate the 30-minute periods characterizing the DS3 services.

Fig. 7 depicts the BESS management during the analyzed days, illustrating the evolution of the power variables $P_{BESS/Exchange}$ (blue line) and $P_{BESS/Reserve}$ (green line) alongside the battery SOC (orange line). Note that positive power values indicate discharging, while negative values denote charging. The displayed results align with the developed environment, as the SOC remains within its limits, and the exchanged power along with reserve power never exceeds the nominal power value. Note that the optimizer is abstaining from exchanging power during the first day and avoiding complete arbitrage cycles during the second day and the morning of the third day, prioritizing a fully charged state to ensure reserve availability for DS3 services while minimizing the battery degradation. However, starting from the fourth day, the optimizer shifts its strategy to perform two complete arbitrage cycles per day, charging and discharging at different intervals to maximize profit from both the DA energy market and the reserve DS3 market.

Fig. 8 illustrates the power management of the PV plant. It shows the evolution of the variables $P_{PV/T}$ (yellow line), $P_{PV/Conv}$ (red line), and $P_{PV/Grid}$ (green line). $P_{PV/Conv}$ is represented as negative to indicate that this power is used to charge the battery, following the sign convention. The figure clearly shows that the optimizer chooses to charge from the PV generation available once a day, rather than buying energy from the DA market. This holds true except for the first two days, where the decision is made not to charge the battery from the PV plant during midday hours, as seen in Fig. 7, where it is also decided not to discharge the battery in the morning. Additionally, on the second day, there is more than 5 MW of PV generation for 3 hours, resulting in the curtailment of $P_{PV/Grid}$ to comply with the inverter limitations, as prescribed by Equation (15).

Fig. 9 is provided to understand this behavior. The upper subplot depicts the evolution of the variables $P_{PV/Conv}$ (yellow), $P_{BESS/EM}$ (blue), and $P_{BESS/Reserve}$ (red). Similarly, the bottom subplot shows the evolution of the DS3 service prices (green) and the DAM prices (orange). The figure reveals how the optimizer dynamically adjusts the battery operation in response to price signals from both the DS3 and DA markets. On the first day, the optimizer refrains from discharging the battery due to

**Table 4**
RES–BESS multi-market participation case study: Profits obtained with the different approaches over the 20 days of testing, categorized into 5-day intervals for each season of the year.

| Season 5 days | Approach | PV Profits | DS3 Profits | DA Profits | Degradation Costs | Total Profits |
|---|---|---|---|---|---|---|
| **Winter** | **PV + DA** | 1463 € | – | 1061 € | 348 € | 2176 € |
| | **PV + DS3** | 1641 € | 814 € | – | – | 2455 € |
| | **DDPG agent** | 1641 € | 701 € | 283 € | 168 € | 2457 € |
| | **TD3 agent** | 1425 € | 647 € | 952 € | 291 € | 2733 € |
| **Spring** | **PV + DA** | 4122 € | – | 785 € | 315 € | 4593 € |
| | **PV + DS3** | 4232 € | 2938 € | – | – | 7170 € |
| | **DDPG agent** | 4140 € | 2300 € | 521 € | 249 € | 6712 € |
| | **TD3 agent** | 4171 € | 2792 € | 465 € | 130 € | 7297 € |
| **Summer** | **PV + DA** | 12,575 € | – | 1511 € | 332 € | 13,754 € |
| | **PV + DS3** | 12,813 € | 814 € | – | – | 13,627 € |
| | **DDPG agent** | 12,595 € | 766 € | 854 € | 117 € | 14,098 € |
| | **TD3 agent** | 12,435 € | 718 € | 1830 € | 256 € | 14,728 € |
| **Autumn** | **PV + DA** | 2634 € | – | 1096 € | 314 € | 3415 € |
| | **PV + DS3** | 2737 € | 2357 € | – | – | 5094 € |
| | **DDPG agent** | 2729 € | 1920 € | 517 € | 263 € | 4903 € |
| | **TD3 agent** | 2669 € | 1885 € | 856 € | 249 € | 5161 € |

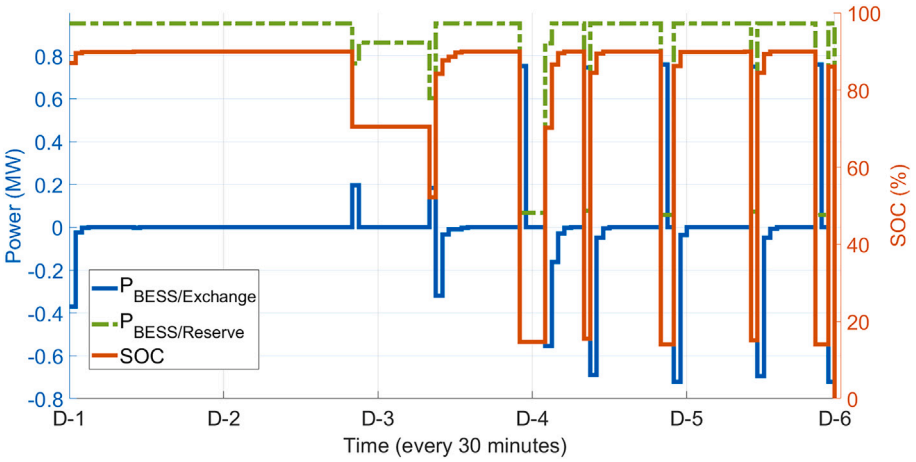



**Fig. 7.** RES–BESS multi-market participation case study: BESS power management for 5 consecutive days.

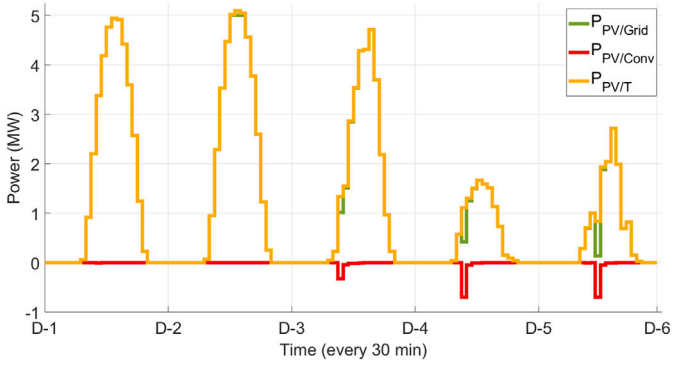



**Fig. 8.** RES–BESS multi-market participation case study: PV power management for 5 consecutive days.

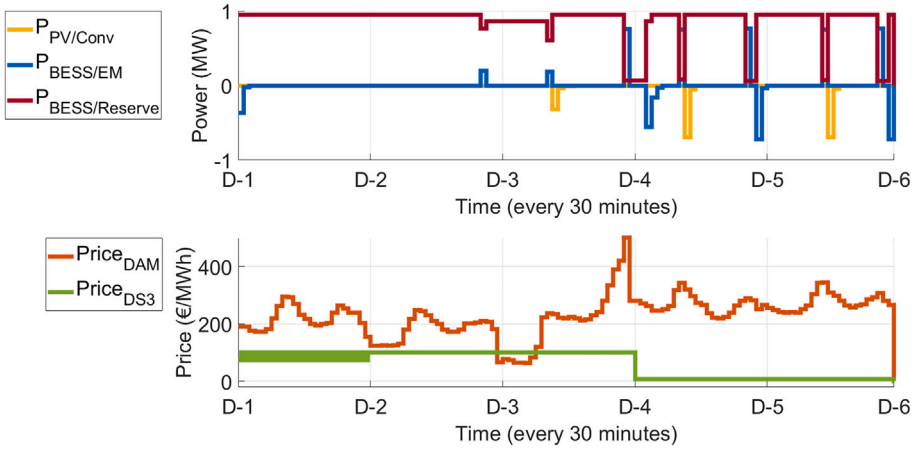



**Fig. 9.** RES–BESS multi-market participation case study: Comparison of the BESS powers and Market prices for 5 consecutive days.

the high DS3 service prices, prioritizing reserve provision over arbitrage. On the second day, it discharges the battery slightly in the afternoon to take advantage of a high DA price spike while still earning revenue from DS3 services. On the third day, the optimizer briefly discharges the battery in the morning and then immediately recharges it from PV generation. Later that night, it detects a peak in the DA price and opts to fully discharge the battery, capitalizing on the price spread. From the fourth day onward, with DS3 prices decreasing, the optimizer switches strategy to perform two full arbitrage cycles per day. It charges the battery from both PV generation and the DA market, aiming to maximize profits by focusing more heavily on DA market opportunities.

In addition, given the data availability, the optimization is extended to the entire dataset of 2022, excluding the training data, to assess if the agent has successfully learned decision-making with the limited training dataset. Table 5 presents the profits obtained for the full year of data using the four approaches. The TD3 agent achieves significantly higher profits compared to the benchmark strategies, with total profits of 709,461€ yielding 8271€ more than the PV + DS3 approach and

**Table 5**
RES–BESS multi-market participation case study: Annual profits.

| | PV profits | DS3 profits | DA profits | Degradation cost | Total Profits |
|---|---|---|---|---|---|
| **PV + DA** | 472,327 € | – | 95,986 € | 25,589 € | 542,724 € |
| **PV + DS3** | 488,424 € | 212,766 € | – | – | 701,190 € |
| **DDPG agent** | 476,814 € | 176,248 € | 60,391 € | 15,361 € | 698,092 € |
| **TD3 agent** | 471,573 € | 173,562 € | 80,935 € | 16,608 € | 709,461 € |

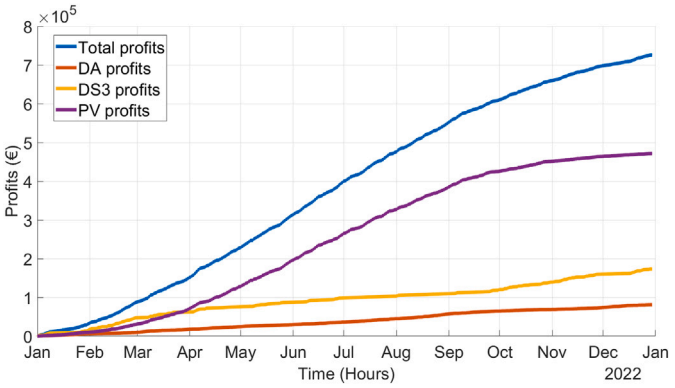



**Fig. 10.** RES–BESS multi-market participation case study: Annual profits evolution.

166,738€ more than the PV + DA approach. Additionally, the TD3 agent outperforms the DDPG agent by 11,369€ . Notably, the TD3 agent maintains higher profits than the PV + DS3 approach, even when degradation costs are considered, a challenge that the DDPG agent cannot overcome. Furthermore, the TD3 agent incurs a degradation cost 8981€ lower than the PV + DA approach, reaffirming the effectiveness of the agent in minimizing the battery degradation. This consistency in results reinforces the robustness of the agent, trained on a limited dataset, and highlights its capability to generalize well. These findings are consistent with the results from the 20-day dataset, further supporting the reliability of this RL strategy. Fig. 10 provides a visual assessment of the profit evolution using the TD3 agent. It illustrates that the main income is derived from the PV Power PPA benefits (purple line), followed by DS3 benefits (yellow line) and DA benefits (orange line). The blue line represents the aggregated sum of the other curves, resulting in final profits of 709,461€.

To sum up, the proposed TD3 agent demonstrates its ability to maximize profits for the RES–BESS installation while complying with the technical constraints of the system. The primary advantage of the proposed method lies in the TD3 agent's adaptability to various stochastic environmental conditions, such as fluctuations in energy market price spreads, reserve AS market price levels, and varying the RES generation. This allows the agent to make dynamic decisions on whether to charge from the grid or RES generation based on market conditions, as well as adjust the number and depth of arbitrages according to the reserve AS market prices.

A significant improvement over the DDPG agent is the TD3 agent's consistent performance across all seasons, particularly in periods with higher variability like spring and autumn. While the DDPG agent struggles to outperform the PV + DS3 benchmark in these scenarios, the TD3 agent consistently achieves higher profits. Additionally, the TD3 agent yields 11,369€ more annual profit compared to the DDPG agent, showcasing its enhanced ability to stabilize the training process and generalize from limited datasets.

Another key advantage of the TD3 agent is its efficiency in training with a limited dataset, thanks to the sequential environment design. Despite the smaller training data, the agent outperforms both benchmark approaches and the DDPG agent. This reduces the computational resources required for training and minimizes the data dependency, making the training process more efficient.

Furthermore, the TD3 agent effectively understands the non-linear battery degradation model, leading to lower degradation costs compared to the PV + DA approach and outperforming the PV + DS3 scenario, which does not consider degradation penalties. This highlights the effectiveness of the proposed non-linear degradation model based on the DOD evolution, which varies adaptively according to the agent's decisions. In addition, this demonstrates the capability of the agent to balance profitability and battery lifetime.

Finally, the TD3 agent offers great versatility as it can be applied to various frameworks, such as different regions with distinct market rules, or to other RES technologies. This flexibility allows the RL agent to adapt to a wide range of case studies, making it highly applicable in different operational contexts.

## 5. Conclusion

This paper introduces a deep RL approach to formulate the bidding strategy of a collocated RES power plant with BESS to participate in multiple day-ahead electricity markets, including energy and ancillary services markets. To achieve this, we develop an MDP framework to define plant operation and address technical constraints while streamlining model training. Furthermore, we include a non-linear battery degradation model that varies depending on the DOD in each time step, which varies adaptively depending on the agent's decisions. As an RL method, we implement an actor–critic method, leveraging a continuous action space, with a focus on the TD3 algorithm, employing two LSTM neural networks as the actor and critic, respectively. The case study is focused on the Irish context, participating in both the day-ahead market and the reserve services for the frequency droop curve response of the DS3 programme. Historical real data from the Irish day-ahead market, SNSP ratio, and PV generation from an actual PV system located in Ireland are considered. This algorithm can be easily implemented in other RES plants, including wind farms, and is adaptable to various electricity markets.

Our findings demonstrate that the RL framework's dynamics adapt to market conditions, leading to significant economic benefits and outperforming benchmark strategies by 8271€/year and 166,738€/year, respectively. In addition, the TD3 agent yields 11,369€ more annual profit compared to the DDPG agent, showcasing its enhanced ability to stabilize the training process and generalize from limited datasets. The RL optimization model exhibits robust behavior in response to variations in PV generation and market conditions across different periods, affirming the effectiveness of the training process. Careful selection of training datasets and simplifications to reduce decision complexity facilitate the model's learning process, minimizing training data requirements and time. Additionally, by incorporating a battery degradation model, we enhance the decision-making process regarding energy storage management, allowing for more accurate assessments of long-term operational costs and performance. Particularly, the RL model consistently achieves lower degradation costs compared to the PV + DA approach, mitigating the effect of the battery degradation. Furthermore, the RL model obtains higher profits than the two benchmark strategies, even though the PV + DS3 approach does not account for any degradation penalties. This highlights the effectiveness of the proposed non-linear degradation model based on the DOD evolution. Moreover, this highlights a unique strength of RL methods: their ability to incorporate complex non-linear degradation models that are challenging to manage with traditional optimization techniques, such as linear programming. This advantage supports the use of RL for informed decision-making in complex and non-linear scenarios, such as the one proposed in this case.

These results have substantial practical implications for grid operators, renewable energy project developers, and market participants,

enhancing project profitability and contributing to more efficient and sustainable grid operation.

Future research will explore the integration of electricity price and energy production predictions to enhance system anticipation and decision-making capabilities further, evaluating the efficiency of the proposed algorithm. To expand the applicability of our findings, we plan to conduct case studies in different contexts. This includes exploring the optimization potential of our approach in wind farm installations, extending beyond solar PV setups. Additionally, we aim to evaluate the effectiveness of our methodology in different electricity market environments, such as the Spanish market, to understand its performance across varied regulatory and operational conditions. Finally, we plan to conduct a comparative analysis among different RL agents such as Proximal Policy Optimization (PPO), and Trust Region Policy Optimization (TRPO), aiming to identify the most effective approach for similar optimization tasks.

## CRediT authorship contribution statement

**Javier Cardo-Miota:** Writing – original draft, visualization, validation, software, methodology, investigation, formal analysis, data curation, and conceptualization. **Hector Beltran:** Writing – review & editing, supervision, resources, project administration, methodology, investigation, and funding acquisition. **Emilio Pérez:** Writing – review & editing, supervision, resources, project administration, investigation, and funding acquisition. **Shafi Khadem:** Writing – review & editing, visualization, validation, supervision, resources, project administration, methodology, investigation, funding acquisition, and conceptualization. **Mohamed Bahloul:** Writing – original draft, visualization, validation, software, methodology, investigation, formal analysis, data curation, and conceptualization.

## Declaration of competing interest

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

## Acknowledgment

The authors would like to acknowledge the support from the following projects and grants: SEAI 21/RDD/690, UJI-B2021-35, PID2021-125634OB-I00, PREDOC/2020/35, and E-2022-26.

## Data availability

Data will be made available upon request.

## References

[1] IRENA. Renewable capacity statistics 2024. Tech. rep.. IRENA; 3 2024.
[2] IEA. The evolution of energy efficiency policy to support clean energy transitions. Tech. rep.. IEA; 2023.
[3] Kroposki B, Johnson B, Zhang Y, Gevorgian V, Denholm P, Hodge B-M, et al. Achieving a 100% renewable grid: Operating electric power systems with extremely high levels of variable renewable energy. IEEE Power Energy Mag 2017;15:61–73. https://doi.org/10.1109/MPE.2016.2637122.
[4] Johnson SC, Rhodes JD, Webber ME. Understanding the impact of non-synchronous wind and solar generation on grid stability and identifying mitigation pathways. Appl Energy 2020;262:114492. https://doi.org/10.1016/j.apenergy.2020.114492.
[5] Hashemi S, Østergaard J. Methods and strategies for overvoltage prevention in low voltage distribution systems with pv. IET Renewable Power Gener 2017;11:205–14. https://doi.org/10.1049/iet-rpg.2016.0277.
[6] Homan S, Brown S. An analysis of frequency events in great britain. Energy Rep 2020;6:63–69. https://doi.org/10.1016/j.egyr.2020.02.028.
[7] Fotis G, Vita V, Maris TI. Risks in the european transmission system and a novel restoration strategy for a power system after a major blackout. Appl Sci 2023;13. https://doi.org/10.3390/app13010083.
[8] Oureilidis K, Malamaki K-N, Gallos K, Tsitsimelis A, Dikaiakos C, Gkavanoudis S, et al. Ancillary services market design in distribution networks: Review and identification of barriers. Energies 2020;13. https://doi.org/10.3390/en13040917.
[9] Fernández-Muñoz D, Pérez-Díaz JI, Guisández I, Chazarra M, Fernández-Espina Á. Fast frequency control ancillary services: An international review. Renew Sustain Energy Rev 2020;120:109662. https://doi.org/10.1016/j.rser.2019.109662.
[10] Groppi D, Pfeifer A, Garcia DA, Krajačić G, Duić N. A review on energy storage and demand side management solutions in smart energy islands. Renewable and Sustain Energy Rev 2021;135:110183. https://doi.org/10.1016/j.rser.2020.110183.
[11] Brivio C, Mandelli S, Merlo M. Battery energy storage system for primary control reserve and energy arbitrage, Sustainable Energy. Grids and Networks 2016;6:152–65. https://doi.org/10.1016/j.segan.2016.03.004.
[12] Ziegler MS, Trancik JE. Re-examining rates of lithium-ion battery technology improvement and cost decline. Energy Environ. Sci. 2021;14:1635–51. https://doi.org/10.1039/D0EE02681F.
[13] Zhang X, Qin CC, Loth E, Xu Y, Zhou X, Chen H. Arbitrage analysis for different energy storage technologies and strategies. Energy Rep 2021;7:8198–206. https://doi.org/10.1016/j.egyr.2021.09.009.
[14] Daggett A, Qadrdan M, Jenkins N. Feasibility of a battery storage system for a renewable energy park operating with price arbitrage. In: 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe); 2017. p. 1–6. https://doi.org/10.1109/ISGTEurope.2017.8260249.
[15] Shin H, Baldick R. Optimal battery energy storage control for multi-service provision using a semidefinite programming-based battery model. IEEE Trans Sustain Energy 2023;14(4):2192–204. https://doi.org/10.1109/TSTE.2023.3263236.
[16] Conte F, Massucco S, Schiapparelli G-P, Silvestro F. Day-ahead and intra-day planning of integrated bess-pv systems providing frequency regulation. IEEE Trans Sustain Energy 2020;11(3):1797–806. https://doi.org/10.1109/TSTE.2019.2941369.
[17] Heredia FJ, Cuadrado MD, Corchero C. On optimal participation in the electricity markets of wind power plants with battery energy storage systems. Computers & Operations Research 2018;96:316–29. https://doi.org/10.1016/j.cor.2018.03.004.
[18] Bahloul M, Daoud M, Khadem SK. Optimal dispatch of battery energy storage for multi-service provision in a collocated pv power plant considering battery ageing. Energy 2024;293:130744. https://doi.org/10.1016/j.energy.2024.130744.
[19] Liu X, Yan Z, Wu J. Optimal coordinated operation of a multi-energy community considering interactions between energy storage and conversion devices. Appl Energy 2019;248:256–73. https://doi.org/10.1016/j.apenergy.2019.04.106.
[20] Li J, Wang W, Yuan Z, Chen J, Zhang Y. Optimal multi-market operation of gravity energy storage and wind power producer using a hybrid stochastic/robust optimization. J Energy Storage 2023;68:107760. https://doi.org/10.1016/j.est.2023.107760.
[21] Namor E, Sossan F, Cherkaoui R, Paolone M. Control of battery storage systems for the simultaneous provision of multiple services. IEEE Trans Smart Grid 2019;10(3):2799–808. https://doi.org/10.1109/TSG.2018.2810781.
[22] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. J Artif Intell Res 1996;4:237–85.
[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. Electric Power Systems Research 2021;196:107229. https://doi.org/10.1016/j.epsr.2021.107229.
[24] Anwar M, Wang C, de Nijs F, Wang H. Proximal policy optimization based reinforcement learning for joint bidding in energy and frequency regulation markets. In: 2022 IEEE Power & Energy Society General Meeting (PESGM); 2022. p. 1–5. https://doi.org/10.1109/PESGM48719.2022.9917082.
[25] Huang B, Wang J. Deep-reinforcement-learning-based capacity scheduling for pv-battery storage system. IEEE Trans Smart Grid 2021;12(3):2272–83. https://doi.org/10.1109/TSG.2020.3047890.
[26] Jeong J, Kim SW, Kim H. Deep reinforcement learning based real-time renewable energy bidding with battery control. IEEE Transactions on Energy Markets, Policy and Regulation 2023;1(2):85–96. https://doi.org/10.1109/TEMPR.2023.3258409.
[27] Cardo-Miota J, Trivedi R, Patra S, Khadem S, Bahloul M. Data-driven approach for day-ahead system non-synchronous penetration forecasting: A comprehensive framework, model development and analysis. Appl Energy 2024;362:123006. https://doi.org/10.1016/j.apenergy.2024.123006.
[28] Denholm PL, Margolis RM, Eichman JD. Evaluating the technical and economic performance of pv plus storage power plants. Tech. rep.. Golden CO (United States): National Renewable Energy Lab.(NREL; 2017.
[29] Baggu MM, Nagarajan A, Cutler D, Olis D, Bialek TO, Symko-Davies M. Coordinated optimization of multiservice dispatch for energy storage systems with degradation model for utility applications. IEEE Trans Sustain Energy 2019;10(2):886–94.
[30] Bahloul M, Khadem SK. Impact of power sharing method on battery life extension in hess for grid ancillary services. IEEE Trans Energy Convers 2019;34(3):1317–27.
[31] Beltran H, Tomás García I, Alfonso-Gil JC, Pérez E. Levelized cost of storage for li-ion batteries used in pv power plants for ramp-rate control. IEEE Trans Energy Convers 2019;34(1):554–61. https://doi.org/10.1109/TEC.2019.2891851.
[32] Li S. Reinforcement Learning for Sequential Decision and Optimal Control. Springer Singapore; 2023. https://doi.org/10.1007/978-981-19-7784-8.
[33] Chen S, Li P, Brady D, Lehman B. Determining the optimum grid-connected photovoltaic inverter size. Sol Energy 2013;87:96–116. https://doi.org/10.1016/j.solener.2012.09.012.
[34] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods, In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning of Proceedings of Machine Learning Research, PMLR; Vol. 80. 2018. p. 1587–96.
[35] Wang H, Miahi E, White M, Machado MC, Abbas Z, Kumaraswamy R, et al. Investigating the properties of neural network representations in reinforcement learning. Artif Intell 2024;330:104100. https://doi.org/10.1016/j.artint.2024.104100.
[36] Wu J, Wu QMJ, Chen S, Pourpanah F, Huang D. A-td3: An adaptive asynchronous twin delayed deep deterministic for continuous action spaces. IEEE Acces 2022;10:128077–89. https://doi.org/10.1109/ACCESS.2022.3226446.

[37] Li K, Ni W, Dressler F. Lstm-characterized deep reinforcement learning for continuous flight control and resource allocation in uav-assisted sensor network. IEEE Internet Things J 2022;9(6):4179–89. https://doi.org/10.1109/JIOT.2021.3102831.
[38] Giorgi M. Buenas expectativas. los precios de baterías de litio perciben una caída del 25% en españa; 2024. https://energiaestrategica.es/caida-de-precios-baterias-espania/.
[39] SEMOpx. report. 3 February 2024 2024.
[40] Bahloul M, Daoud M, Khadem SK. A bottom-up approach for techno-economic analysis of battery energy storage system for irish grid ds3 service provision. Energy 2022;245:123229. https://doi.org/10.1016/j.energy.2022.123229.
[41] EirGrid. SONI, Ds3 system services protocol regulated arrangements ds3 system services implementation projecte1st may 2019 version 2.0.. Tech. rep.. EirGrid-SONI; 9 2019.
[42] IERC. Energy storage in ireland: Barriers and policy interventions. Tech. rep.. IERC; 5 2021.
[43] EirGrid. SONI, Ds3 system services: Statement od payments. applicable from 01st january 2022, Tech. rep.. EirGrid-SONI; 12 2021.
[44] EirGrid. System Non-Synchronous Penetration Definition and Formulation. Tech. Rep. August. EirGrid; 2018.
[45] EirGrid Group. Smart grid dashboard. 2024. p. 2024–01–10. https://www.smartgriddashboard.com/, accessed:.
[46] SEMOpx. Market data. 2024. p. 2024–01–10. https://www.semopx.com/market-data/, accessed:.