Simulation of Microgrid Energy Management under Battery Degradation Costs: a PPO-Based Reinforcement Learning Approach

Gianluca Ferro*, Alessio Orlandi*, Francesco Giuseppe Quilici*, Giovanni Lutzemberger*, Enrico De Santis*, and Antonello Rizzi*
*Department of Information Engineering, Electronics and Telecommunications
University of Rome "La Sapienza"
Via Eudossiana 18, 00184 Rome, Italy
Email: gianluca.ferro@uniroma1.it, orlandi.1954535@studenti.uniroma1.it, francesco_giuseppe.quilici@phd.unipi.it, giovanni.lutzemberger@unipi.it, enrico.desantis@uniroma1.it, antonello.rizzi@uniroma1.it

***Abstract*—Residential microgrids with photovoltaic generation and battery storage require energy management strategies that reduce grid costs while limiting long-term battery degradation. We present a degradation-aware simulator in which the storage system is managed by a Battery Management System module based on an equivalent circuit model (ECM) with SoH-dependent parameters. Aging is updated from experimentally identified SoH-Ah throughput curves for a commercial NMC 18650 cell, and a DoD-based wear cost is included in the objective. A PPO controller is trained to select charge/discharge setpoints and is tested on real household data against a rule-based controller and an oracle MPC benchmark. The best policies consistently outperform the rule-based baseline and approach oracle MPC performance at medium battery utilization.**

***Index Terms*—microgrid, energy management, battery degradation, reinforcement learning, model predictive control**

I. Introduction

Residential microgrids with PV and battery storage require EMS strategies that reduce grid costs while limiting battery aging. Existing degradation-aware EMS methods often rely on simplified battery models with fixed efficiencies and weakly chemistry-grounded aging assumptions, which can misestimate the economic impact of capacity fade and resistance growth.

In this study, we propose a degradation-aware simulation framework in which a dedicated Battery Management System (BMS) module implements an equivalent circuit model (ECM) with State of Health (SoH) dependent parameters. Aging is updated from experimentally identified SoH-Ah throughput curves, and a depth-of-discharge (DoD) wear cost is included in the objective. On top of this simulator, a Proximal Policy Optimization (PPO) controller learns battery charge/discharge setpoints and is evaluated on real household data against a rule-based baseline and an oracle Model Predictive Control (MPC) benchmark.

Experimental results demonstrate that the learned policy consistently outperforms standard rule-based strategies and achieves a cost-degradation trade-off comparable to the oracle MPC in medium-utilization regimes, effectively balancing immediate economic gains with long-term asset preservation. The main contributions of this work can be summarized as follows:

- We develop a degradation-aware simulation environment for residential microgrid energy management in which the EMS is coupled with a battery module based on an ECM with State of Health dependent parameters.
- We incorporate battery lifetime effects into the EMS through experimentally derived aging curves and a DoD-based wear cost, so that control decisions account for the long term economic impact of battery cycling.
- We train a PPO-based controller to generate battery charge/discharge setpoints and benchmark it against a rule-based EMS and an oracle MPC on real household demand and PV data, analyzing the resulting policies in the cost/degradation Pareto plane.

Data used in this work is available at [1]. The code associated with this study will not be made publicly available at this stage, as it is part of an ongoing research project currently under active development. Public release is planned in a subsequent phase, once the research framework has been fully consolidated.

The paper is organized as follows.
Section II covers related work while the methodology, including the simulator, benchmarks, and RL agent, is presented in Secs. III–V. Finally, Sec. VI details the battery characterization, with results and conclusions in Secs. VII–IX.

II. Related Work

Data-driven approaches are increasingly used in smart grid systems, from operational optimization to fault detection and classification [2]. Optimization-based approaches such as Mixed-Integer Linear Programming (MILP), Model Predictive Control, and stochastic MPC are widely used for microgrid EMS design. They minimize operating cost by optimizing grid import/export while enforcing power-flow constraints, SoC

dynamics and bounds, and power balance at each step [3]. Degradation is increasingly included by adding an equivalent wear cost, often derived from depth-of-discharge or cycle-counting models [4]. However, most formulations still use simplified batteries, typically State of Charge (SoC) dynamics with constant efficiency.

Reinforcement learning (RL) has emerged as an alternative to optimization-based EMS, motivated by its ability to handle nonlinear dynamics, uncertainty, and complex decision processes without requiring explicit system models. Recent works apply, for example, deep RL algorithms such as DQN, DDPG, TD3, SAC, and PPO to microgrid energy management problems, often reporting lower operating costs compared to rule-based controllers or heuristic baselines [5]. Among policy-gradient methods, Proximal Policy Optimization (PPO) and its multi-agent variants are increasingly adopted for EMS and ESS scheduling, owing to their stable updates and suitability for continuous control [6]. In these studies, battery aging is not commonly introduced as an additional internal health state with long-term dynamics; instead, when lifetime considerations are included, they are typically addressed through surrogate terms, for instance by discouraging frequent charge and discharge switching.

Despite the critical importance of battery storage in microgrids, most EMS formulations adopt simplified models based on SoC dynamics and constant charge-discharge efficiencies. In contrast, equivalent circuit models (ECM), widely used in BMSs to capture voltage behavior, internal resistance, and losses, are only sporadically integrated into EMS frameworks due to their increased computational complexity and parameter dependence on operating conditions. From an aging perspective, extensive experimental literature exists on battery degradation, where capacity fade is often characterized as a function of charge throughput or cycling history [7]. However, only a limited number of EMS or RL-based studies directly integrate experimentally derived aging curves into the control or learning process.

III. System Modeling and Problem Formulation

We simulate a microgrid comprising an aggregated load, a photovoltaic generator, a battery energy storage system (BESS), and a point of common coupling (PCC) enabling bidirectional exchange with the main grid. Importantly, the simulator represents the physical layer, while the Energy Management System is the decision layer that schedules energy flows. In this work, we focus on tertiary EMS (economic/strategic optimization) while secondary and primary control (voltage/frequency and local device control) are assumed to be handled by the underlying controllers. The overall microgrid architecture and EMS decision layer are summarized in Fig. 1.

A. Microgrid Simulator

The microgrid is simulated in discrete time using `pymgrid` [8]. Load and PV generation are driven by exogenous time series, while the grid module models bidirectional

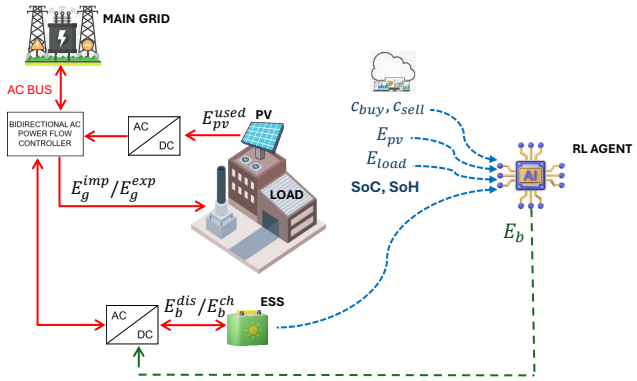

Fig. 1. Scheme of the simulated microgrid and EMS decision layer. Energy flows are represented by red lines, while EMS input data flows are the blue dashed lines. The green dashed line is the EMS battery action.

exchange at the PCC within import/export limits; imbalance due to constraints is handled via loss of load and curtailment terms. For the storage system, we implement a custom *battery module* that acts as a BMS. Specifically, it receives a charge/discharge setpoint from the EMS, checks admissibility, and applies the corresponding internal command.

The battery evolution is captured by a chemistry-dependent transition model. Open-circuit voltage and internal resistance are taken from experimental lookup tables versus SoC and SoH. Hence, given the energy setpoint, current and terminal voltage are computed through the ECM (including ohmic losses) and used to update energy/charge states (SoE/SoC). Aging is updated from experimentally identified cumulative Ah throughput curves and reduces effective capacity, with limits updated accordingly.

A detailed description of the battery pack characteristics and of the experiments conducted to obtain cell level data is provided in Section VI.

B. Operational Constraints

The simulation is performed in discrete timesteps. At time step k, the available photovoltaic production and the demand are exogenous signals. The EMS acts on the battery and grid exchange setpoints, which the simulator interprets as net energy flows over the interval. We introduce the non-negative variables $E_{\mathrm{b}}^{\mathrm{ch}}(k)$ and $E_{\mathrm{b}}^{\mathrm{dis}}(k)$ for charge and discharge, and $E_{\mathrm{g}}^{\mathrm{imp}}(k)$ and $E_{\mathrm{g}}^{\mathrm{exp}}(k)$ for import and export, assuming that in each interval the command is unidirectional, so that charge and discharge, and import and export, do not occur simultaneously.

The energy balance at the PCC at each step can therefore be expressed as:

$$E_{\mathrm{pv}}^{\mathrm{used}}(k)+E_{\mathrm{g}}^{\mathrm{imp}}(k)+E_{\mathrm{b}}^{\mathrm{dis}}(k) = E_{\mathrm{load}}^{\mathrm{met}}(k)+E_{\mathrm{g}}^{\mathrm{exp}}(k)+E_{\mathrm{b}}^{\mathrm{ch}}(k). \quad (1)$$

TABLE I
TIME OF USE TARIFFS, TOU, USED FOR GRID EXCHANGE.

Band	Time window, local	c_{buy}	c_{sell}
Peak	18 to 20	0.35	0.123
Standard	7 to 17; 21 to 22	0.30	0.123
Off peak	Remaining hours	0.27	0.123

Grid exchange is limited by connection capacity constraints, given by:

$$0 \le E_{\mathrm{g}}^{\mathrm{imp}}(k) \le \overline{E}_{\mathrm{imp}}, \tag{2}$$

$$0 \le E_{\mathrm{g}}^{\mathrm{exp}}(k) \le \overline{E}_{\mathrm{exp}}, \tag{3}$$

where $\overline{E}_{\mathrm{imp}}$ and $\overline{E}_{\mathrm{exp}}$ are the maximum energies that can be exchanged per interval in import and export.

For the storage system, the BMS enforces operational constraints on both the internal state and the admissible energy flows. Degradation is represented through the state of health $\mathrm{SoH}(k)$, which reduces the usable pack capacity. In particular, letting C_{nom}^{E} denote the nominal energy capacity, in kWh, the effective energy capacity is:

$$C_{\mathrm{eff}}^{E}(k) = \mathrm{SoH}(k)\, C_{\mathrm{nom}}^{E}. \tag{4}$$

Based on $E(k)$, the energy stored in the BESS, we define the *State of Energy*, SoE, as a normalized quantity:

$$\mathrm{SoE}(k) = \frac{E(k)}{C_{\mathrm{eff}}^{E}(k)}, \tag{5}$$

subject to the constraints:

$$\mathrm{SoE}_{\min} \le \mathrm{SoE}(k) \le \mathrm{SoE}_{\max}. \tag{6}$$

Similarly, introducing the pack charge $Q(k)$, in Ah, and the effectively available charge $Q_{\mathrm{eff}}(k)$, in Ah, we define the *State of Charge*, SoC:

$$\mathrm{SoC}(k) = \frac{Q(k)}{Q_{\mathrm{eff}}(k)}, \tag{7}$$

with the corresponding constraints

$$\mathrm{SoC}_{\min} \le \mathrm{SoC}(k) \le \mathrm{SoC}_{\max}. \tag{8}$$

C. Microgrid Operational Costs

The evaluation of the energy management strategies is carried out in terms of operating cost, including the costs and revenues associated with energy exchange with the main grid at the PCC, and a storage degradation cost introduced to monetize the wear associated with charge and discharge cycles.

1) Grid exchange costs: The purchase and selling tariff signals $c_{\mathrm{buy}}(k)$ and $c_{\mathrm{sell}}(k)$ are assumed known and are expressed in $\mathrm{EUR/kWh}$. In this study, we adopt *Time of Use* tariffs, denoted as ToU, assigned according to the hour of day. The ToU bands used in the simulations are reported in Tab. I. The grid exchange cost over a single interval is then:

$$\begin{aligned} J_{\mathrm{grid}}(k) = {} & c_{\mathrm{buy}}(k)\, E_{\mathrm{g}}^{\mathrm{imp}}(k) \\ & - c_{\mathrm{sell}}(k)\, E_{\mathrm{g}}^{\mathrm{exp}}(k), \end{aligned} \tag{9}$$

which is positive in case of net import and can become negative, representing revenue, when export is remunerated.

2) Battery wear cost model: To include the effect of aging in the EMS objective function in a computationally efficient manner, we adopt the wear cost model proposed in [9], which links the degradation cost to depth of discharge. Defining

$$\mathrm{DoD}(k) = 1 - \mathrm{SoC}(k), \tag{10}$$

the model introduces a degradation acceleration curve, denoted as ACC, as a function of DoD:

$$\mathrm{ACC}(\mathrm{DoD}) = a_0\, \mathrm{DoD}^{-b_0} \exp(-c_0\, \mathrm{DoD}), \tag{11}$$

and defines an incremental wear cost at step k proportional to the variation of the quantity $1/\mathrm{ACC}$ between two successive instants:

$$\begin{aligned} C_{\mathrm{wear}}(k) = \alpha\, \frac{B_{\mathrm{rep}}}{2} \Big| & \frac{1}{\mathrm{ACC}(\mathrm{DoD}(k))} \\ & - \frac{1}{\mathrm{ACC}(\mathrm{DoD}(k-1))} \Big|, \end{aligned} \tag{12}$$

where B_{rep} represents the replacement cost of the battery pack and α is a multiplicative factor, possibly temperature dependent, that allows modulating the economic impact of degradation.

The parameters a_0, b_0, and c_0 in (11) were obtained by fitting the semi-empirical DoD-dependent cycle-life law in [9] to experimental cycle-life data for NMC 18650 cells reported by Ecker et al. [10]. Specifically, the discrete pairs $(\mathrm{DoD}, \mathrm{ACC})$ were reconstructed from the reported number of equivalent full cycles to end-of-life at each DoD level, and the three parameters were identified via nonlinear least-squares regression. The resulting values used in the simulations are $a_0 = 1354$, $b_0 = 1.614$, and $c_0 = 0.068$.

This model was selected because it allows associating to wear a marginal cost linked to depth of discharge, naturally penalizing deeper cycles, and because it is compatible with integration into EMS optimization problems.

IV. EMS BENCHMARK MODELS

A. Rule Based EMS

As baseline, a rule-based controller (RBC) was adopted, that uses only instantaneous demand $E_{\mathrm{load}}(k)$ and PV production $E_{\mathrm{pv}}(k)$, without solving an optimization problem. Defining the net load $NL(k) = E_{\mathrm{load}}(k) - E_{\mathrm{pv}}(k)$, the policy prioritizes self-consumption: i) if $NL(k) > 0$, it discharges the battery up to $\overline{E}_{\mathrm{b}}^{\mathrm{dis}}(k)$ and imports the residual from the grid; ii) if $NL(k) < 0$, it charges the battery up to $\overline{E}_{\mathrm{b}}^{\mathrm{ch}}(k)$ and exports the residual. Admissibility and mutual exclusion are enforced by the BMS, and the resulting exchanges satisfy the balance (1).

B. Model Predictive Control

The EMS based on MPC determines the exchange setpoints by optimizing, at each interval k, a sequence of decisions over a finite prediction horizon N, and then applying only the first decision according to a *receding horizon* scheme. In this work,

an *oracle* MPC is considered, since the future horizon values of demand $E_{\mathrm{load}}(k+i)$ and photovoltaic production $E_{\mathrm{pv}}(k+i)$, are given as input to the model. This assumption allows using the MPC controller as a performance upper bound in the comparison with other EMS strategies.

1) Decision variables and model: At the generic predicted instant $k+i$, with $i = 0, \dots, N-1$, the MPC optimizes the BESS charge and discharge energies $E_{\mathrm{b}}^{\mathrm{ch}}(k+i)$ and $E_{\mathrm{b}}^{\mathrm{dis}}(k+i)$, and the grid import and export energies $E_{\mathrm{g}}^{\mathrm{imp}}(k+i)$ and $E_{\mathrm{g}}^{\mathrm{exp}}(k+i)$, all non negative.

Due to the MILP nature of MPC, electrochemical nonlinearities are not included directly in the prediction model. In particular, BESS losses are represented through a *dynamic* efficiency $\eta_{\mathrm{dyn}}(k)$, updated at each step from the ECM parameters and treated as a known parameter in the optimization. The BESS dynamics are therefore formulated in terms of stored energy $E(k)$, in kWh, as:

$$E(k+i+1) = E(k+i) + \eta_{\mathrm{dyn}}(k+i)\, E_{\mathrm{b}}^{\mathrm{ch}}(k+i) - \frac{1}{\eta_{\mathrm{dyn}}(k+i)}\, E_{\mathrm{b}}^{\mathrm{dis}}(k+i). \tag{13}$$

In the simulator, $\eta_{\mathrm{dyn}}(k)$ is defined from the adopted ECM in Sec. VI through the open circuit voltage V_{OC} and the ohmic resistance R_0. Letting the terminal voltage be:

$$V(k) = V_{OC}(k) - R_0(k)\, I(k), \tag{14}$$

with $I(k)$ positive during discharge, we adopt a step level efficiency, including a constant inverter factor, equal to the ratio between useful power and electrochemical power:

$$\eta_{\mathrm{dyn}}(k) = \eta_{\mathrm{inv}} \begin{cases} \dfrac{V(k)}{V_{OC}(k)}, & I(k) \geq 0, \\ \dfrac{V_{OC}(k)}{V(k)}, & I(k) < 0, \end{cases} \qquad \eta_{\mathrm{dyn}}(k) \in [0,1]. \tag{15}$$

where $\eta_{\mathrm{inv}} \in (0,1]$ accounts for the power electronics losses and η_{dyn} is saturated in $[0,1]$ for numerical robustness. The BESS energy limits derive from the constraints on SoE and from the effective capacity C_{eff}^{E}, which depends on SoH (see Sec. III):

$$\mathrm{SoE}_{\min}\, C_{\mathrm{eff}}^{E}(k) \leq E(k+i) \leq \mathrm{SoE}_{\max}\, C_{\mathrm{eff}}^{E}(k), \qquad i = 0, \dots, N. \tag{16}$$

Finally, for each $i = 0, \dots, N-1$, the energy balance at the PCC (1) must hold.

2) Optimization problem: The MPC computes the command sequence by minimizing a cumulative cost that combines the energy exchange cost with the grid, with $c_{\mathrm{buy}}(k)$ and $c_{\mathrm{sell}}(k)$, penalties for curtailment and loss of load, and a BESS degradation term derived from the wear cost model in Sec. III-C. A formulation compatible with the simulator is:

$$\min \sum_{i=0}^{N-1} \Big(c_{\mathrm{buy}}(k+i)\, E_{\mathrm{g}}^{\mathrm{imp}}(k+i) - c_{\mathrm{sell}}(k+i)\, E_{\mathrm{g}}^{\mathrm{exp}}(k+i) + \lambda_{\mathrm{ll}} \big(E_{\mathrm{load}}(k+i) - E_{\mathrm{load}}^{\mathrm{met}}(k+i) \big) + \lambda_{\mathrm{curt}} \big(E_{\mathrm{pv}}(k+i) - E_{\mathrm{pv}}^{\mathrm{used}}(k+i) \big) \Big) + \lambda_{\mathrm{wear}}\, J_{\mathrm{deg}}. \tag{17}$$

The weights λ_{ll} and λ_{curt} are introduced in order to discourage, respectively, loss of load and curtailment with respect to grid exchange, while ensuring problem feasibility under tight constraints. The scalar λ_{wear} is the *wear cost scale* used to tune the relative importance of the degradation term in MPC.

The term J_{deg} is defined in Sec. III-C starting from the quantity $\phi(\mathrm{SoC}) = 1/\mathrm{ACC}(1-\mathrm{SoC})$. To keep its inclusion in a MILP problem, $\phi(\mathrm{SoC})$ is approximated by a piecewise affine function with 3 segments over the operating interval $[\mathrm{SoC}_{\min}, \mathrm{SoC}_{\max}]$.

V. Reinforcement Learning Agent

The proposed EMS is formulated as a *reinforcement learning* problem in which a parametric policy learns to determine the storage management setpoints in order to minimize the microgrid operating cost, including the economic impact of battery degradation. The agent is trained using the Proximal Policy Optimization algorithm (PPO), implemented through the `Stable-Baselines3` library [11]. The RL agent does not receive load and PV production forecasts as inputs, since the objective of this work is to study the ability of RL to learn policies that reduce wear costs, while isolating the agent decisions from the possible influence of uncertainty associated with external forecasting models. In this way, performance is evaluated in a more robust and reproducible setting, which does not depend on forecast quality, and keeping the observation dimensionality limited.

A. PPO Algorithm

Proximal Policy Optimization (PPO) is an on-policy actor–critic method that promotes stable learning via conservative updates, while keeping first-order gradient optimization. The control policy is a parameterized stochastic policy $\pi_\theta(a \mid s)$ mapping the observable state s to a distribution over actions a. In parallel, a value function $V_\phi(s)$ estimates the expected return and reduces the variance of policy-gradient estimates. Moreover, starting from trajectories collected by rolling out the current policy $\pi_{\theta_{\mathrm{old}}}$, PPO updates θ by maximizing a surrogate objective built on importance sampling. Defining the probability ratio:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)}, \tag{18}$$

and an advantage estimate $\hat{A}_t$ (computed from the observed rewards and a baseline V_ϕ, e.g., via generalized advantage estimation), the clipped PPO objective is:

$$L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t\big[\min\big(r_t(\theta)\,\hat{A}_t, \\ \mathrm{clip}\big(r_t(\theta), 1-\varepsilon, 1+\varepsilon\big)\,\hat{A}_t\big)\big], \tag{19}$$

where ε limits the effective change of the policy between successive iterations. The clipping mechanism discourages overly large updates and improves training stability by preventing harsh policy updates [12].

PPO was selected in this work because it naturally addresses the continuous EMS action, i.e., the storage energy setpoint, without resorting to discretization. Moreover, the on-policy nature of PPO is well aligned with the present setting, where the reward explicitly internalizes a degradation-related cost and the state includes a slowly varying component, namely the SoH, so that stable updates can help mitigate training instabilities associated with long-horizon effects.

B. RL EMS Formulation

a) Environment and simulator interface: The RL environment is implemented using `Gymnasium` and encapsulates the microgrid simulator described in Sec. III-A. At each interval k, demand $E_{\mathrm{load}}(k)$ and photovoltaic production $E_{\mathrm{pv}}(k)$ are read from the dataset together with the tariff signal $(c_{\mathrm{buy}}(k), c_{\mathrm{sell}}(k))$, and are used to build the observation provided to the policy. The agent then returns a charge or discharge setpoint for the BESS. Based on this request, the environment determines the grid exchange consistent with the energy balance and advances the simulation by one step.

b) Observations: The observable state is constructed from the variables available at the tertiary EMS level. In the adopted configuration, the observation includes demand $E_{\mathrm{load}}(k)$, production $E_{\mathrm{pv}}(k)$, the internal storage state through $\mathrm{SoC}(k)$ and $\mathrm{SoH}(k)$, and the instantaneous limits on battery exchangeable energy $\overline{E}_{\mathrm{b}}^{\mathrm{dis}}(k)$ and $\overline{E}_{\mathrm{b}}^{\mathrm{ch}}(k)$. To allow the policy to capture intraday seasonality and price dependence, cyclic temporal features, hour and day of week, and the ToU tariffs $(c_{\mathrm{buy}}(k), c_{\mathrm{sell}}(k))$ are also included.

c) Actions: The agent produces a scalar continuous action that represents the signed BESS energy setpoint over a single interval, interpreted as a discharge request (positive) or a charge request (negative), which are mutually exclusive. The command is made admissible by the BMS through saturation within the limits $\overline{E}_{\mathrm{b}}^{\mathrm{dis}}(k)$ and $\overline{E}_{\mathrm{b}}^{\mathrm{ch}}(k)$. Once the battery exchange is fixed, the grid exchange $\big(E_{\mathrm{g}}^{\mathrm{imp}}(k), E_{\mathrm{g}}^{\mathrm{exp}}(k)\big)$ is determined automatically as the residual required by the balance equation (1), while respecting the constraints (3).

d) Reward: The reward is defined as the negative of a weighted instantaneous cost. The main terms reflect the grid exchange cost $J_{\mathrm{grid}}(k)$ (9) and the wear cost $C_{\mathrm{wear}}(k)$ (12). An action violation term $v_{act}(k)$ is also added, which is a boolean flag that is True when the requested action is not feasible in terms of battery charge and discharge limits. In addition, a penalty term $v_\mu(k)$ for excessive micro charge and discharge throughput is included, which is equal to the instantaneous battery energy throughput when this is positive but below a predefined threshold E_μ, and zero otherwise.

In compact form, the reward can then be defined as follows:

$$r(k) = -\,w_{\mathrm{eco}}\,J_{\mathrm{grid}}(k) - w_{\mathrm{wear}}\,C_{\mathrm{wear}}(k) \\ - w_{\mathrm{act}}\,v_{\mathrm{act}}(k) - w_\mu\,v_\mu(k). \tag{20}$$

VI. Energy Storage System

A. ECM model adopted

The battery energy storage system is modeled with a low order ECM, namely a zero-RC (Rint) model, chosen for low computational burden in long horizon simulations [13]. This choice is consistent with quarter-hourly data for which higher order battery models would not be justified, as additional dynamic states would not be properly exploited at this time scale. The model is purely electrical, and no thermal dynamics are simulated. Temperature enters only through temperature dependent parameters treated as exogenous inputs. All simulations are performed at ambient temperature, $T = 20°\mathrm{C}$.

The adopted ECM consists of an open-circuit voltage source V_{OC} in series with an ohmic resistance R_0.

The terminal voltage of the battery pack is computed as:

$$V = V_{OC}(SoC, T, SoH) - R_0(SoC, T, SoH)\cdot I, \tag{21}$$

where I is the battery current (positive during discharge). Both V_{OC} and R_0 are parameterized as functions of SoC, temperature T, and SoH, using experimentally derived lookup tables and trilinear interpolation, following the approach presented in [13]. Cell-level parameters are scaled to pack level as:

$$V_{OC} = V_{OC_{cell}} \cdot N_s \tag{22}$$

$$R_0 = R_{0_{cell}} \cdot \frac{N_s}{N_p}, \tag{23}$$

where N_s and N_p are respectively the number of cells in series and in parallel.

Aging is modeled separately from the electrical dynamics via a throughput-based law. The cumulative absolute Ah throughput is tracked at cell level, and the corresponding SoH is updated by interpolating the experimental SOH-Ah curves. This approach guarantees reliable estimates of SoH, which affects both effective capacity and electrical parameters.

B. Cell characterization

The experimental characterization was carried out using a Chroma 17020 battery cycler equipped with eight channels, each one rated at 100 V and 50 A, and an FDM1000 climatic chamber (-20°C to 80°C). The tested cell is a commercial cylindrical NMC cell, whose specifications are shown in Table II.

The cell was subjected to a realistic cycling profile derived from automotive applications. The adopted profile results in an

TABLE II
CELL SPECIFICATIONS.

Cell typology EVE-INR 18650/33V	Lithium-ion NMC
Geometry	Cylindrical
Nominal capacity $\mathbf{Q}_{\mathrm{n,cell}}$ (Ah)	3.2
Nominal energy $\mathbf{E}_{\mathrm{n,cell}}$ (Wh)	11.5
Voltage range: max–min–nominal (V)	4.2 – 2.5 – 3.6
Max continuous discharge (A)	10
Operating temperature (°C)	Charge: $0°\mathrm{C}-50°\mathrm{C}$ Discharge: $-20°\mathrm{C}-60°\mathrm{C}$

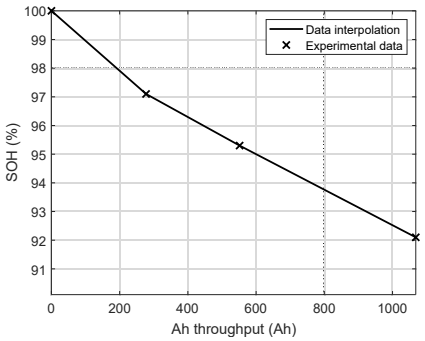

Fig. 2. State of Health trend resulting from cell aging test in function of the cumulative capacity discharged.

average discharge C-rate of approximately 1C and an average charge C-rate of about 0.5C. Consequently, the electrical stress imposed on the cell is moderate and representative of typical operating conditions expected in residential stationary storage systems [14].

Fig. 2 shows the experimentally obtained SoH-Ah throughput relationship used to model battery aging, obtained by performing 400 charging-discharging cycles.

Fig. 3 reports the corresponding evolution of the electrical parameters V_{OC} and R_0 as a function of SoC, represented for different SoH levels and evaluated at a reference temperature of 20°C.

While the open-circuit voltage exhibits only marginal variations with aging over the considered SoH range, the ohmic resistance shows a clear and systematic increase as SoH decreases. These trends confirm that battery aging predominantly affects the resistive behaviour of the cell.

C. Battery sizing

The battery pack is configured with N_s series cells and N_p parallel strings ($N_{cells} = N_s \cdot N_p$). The choice of N_s is constrained by typical residential inverter voltage ranges (e.g. 120-550 V). Accordingly, N_s is chosen such that the nominal pack voltage is in the middle range, ensuring that the minimum battery voltage, when discharged, is not lower than the minimum voltage allowed by the inverter, yielding:

$$N_s = \frac{V_{n_{pack}}}{V_{n_{cell}}}. \tag{24}$$

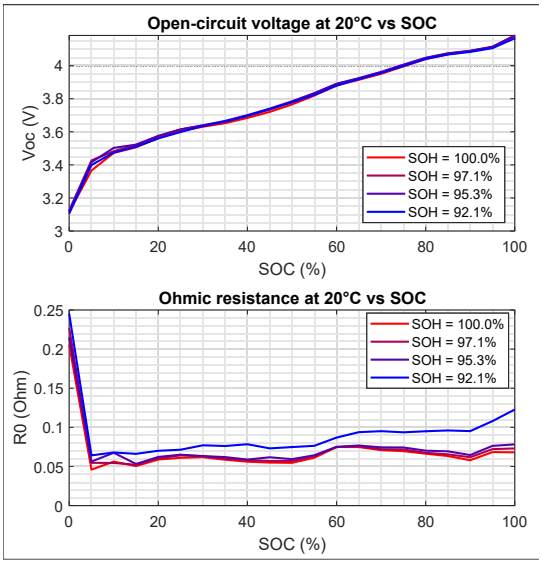

Fig. 3. Parameters evolution with aging. The values are updated within the battery model according to the state of health of the battery at each time step.

Once N_s is defined, the number of parallel strings N_p is selected to meet the target energy E_{target} according to:

$$N_p = \frac{E_{target}}{N_s \cdot V_{n_{cell}} \cdot Q_{n_{cell}}}. \tag{25}$$

For the considered scenario, a target energy of $E_{target} = 17.5$ kWh is adopted, resulting in a configuration with $N_s = 87$ and $N_p = 17$, for a total of $N_{cells} = 1479$.

VII. EXPERIMENTAL SETUP

A. Dataset

The load and photovoltaic production profiles used in the simulator are derived from Open Power System Data (OPSD) [1]. This dataset provides load and PV generation at 15 minutes resolution from small commercial users and residential households.

Starting from the available appliance measurements, an aggregated demand profile was obtained by summing all device consumption series. The training set spans 16 months from September 2015 to December 2016, while the test set spans 2 months from January 2017 to March 2017.

B. RL Model Training Settings

The RL agent is trained offline on the training set previously described, using a discrete-time simulator with sampling time $T_s = 15$ min. Each training episode spans one month, i.e., $N_{\mathrm{ep}} = 2880$ steps. To increase the diversity of operating conditions seen during learning, each episode starts from a randomly selected index in the training time series and the battery initial state of charge is randomized uniformly within the admissible interval $[\mathrm{SoC}_{\min}, \mathrm{SoC}_{\max}]$. Training is carried out with a single environment, and $N_{\mathrm{train}} = 200$ episodes, corresponding to $N_{\mathrm{train}} \cdot N_{\mathrm{ep}} = 576{,}000$ interaction steps.

The agent interface (observations, action, and reward structure) follows the RL EMS formulation in Sec. V-B.

TABLE III
PPO TRAINING HYPERPARAMETERS USED IN THE CASE STUDY.

Parameter	Value
Policy network (actor/critic)	$[64, 64]$
Learning rate	3×10^{-4}
γ	0.99
GAE λ	0.95
Clip range ε	0.20
n_{steps}	512
Batch size	128
Epochs per update	10
Entropy coefficient	3×10^{-3}

Regarding the reward composition described in Sec. V-B, the weights used in the case study are set to $w_{\text{eco}} = 2.3$ (economic term), and $w_{\mu} = 0.7$ (micro-throughput penalty with $E_{\mu} = 0.15$).

In order to evaluate the RL model under different wear cost penalization settings, a training sweep was done with the following values of w_{wear} : [0.2, 0.3, 0.4, 0.5, 0.6].

C. Simulations

All EMS strategies are evaluated on the test set using the same microgrid parameters and tariff structure described in the previous sections. The proposed PPO-based EMS (Sec. V) was compared against the RBC baseline (Sec. IV-A) and the oracle MPC benchmark (Sec. IV-B). In addition, the MPC benchmark was evaluated by running multiple MPC tests with different values of the wear cost scale λ_{wear}, namely $[0.1, 0.3, 0.4, 0.5]$, in order to explore the trade-off between economic performance and degradation.

Performance was assessed by reporting cumulative import cost, export revenue, and net grid cost, together with the battery wear cost and final SoH value.

Moreover, we compare RL and MPC through the empirical Pareto front in the objective space (battery wear cost, net grid cost), to visualize the trade-off between operating cost and degradation.

VIII. ANALYSIS OF RESULTS

This section reports the performance of the considered EMS strategies on the test set, focusing on the trade-off between grid cost reduction and battery aging. Specifically, Fig. 4 summarizes the Pareto analysis by plotting, for each simulated operation, the net grid economic cost on the y-axis and the battery wear cost on the x-axis.

Table IV reports a representative subset of MPC and RL policies (together with the RBC baseline) on the test horizon, decomposing the total return into economic and battery-wear contributions. From the table, it can be observed that the best RL configurations achieve a total cost that is close to the MPC benchmark and consistently improve upon the RBC baseline. Compared to MPC, which requires solving an optimization problem online at every control step, a trained RL policy only performs a single neural network forward pass. As a result, inference latency is low and nearly constant, making the RL

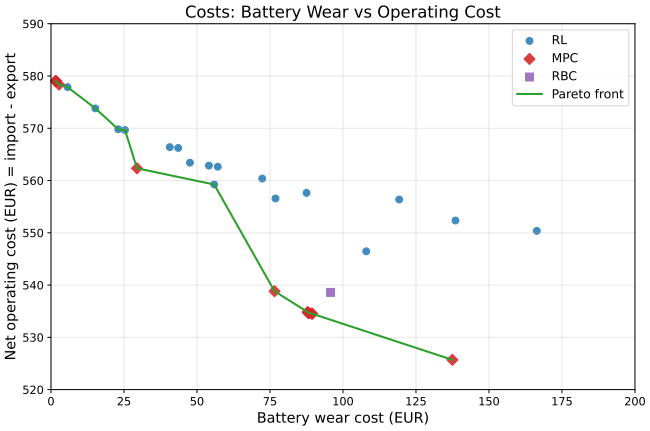

Fig. 4. Pareto front of net grid economic cost (y-axis) versus battery wear cost (x-axis) for the considered EMS strategies.

TABLE IV
COMPARISON OF FINAL COSTS AND SOH ON THE TEST HORIZON.

Model	R_{eco} [€]	C_{wear} [€]	R_{tot} [€]	$\text{SoH}_{\text{final}}$ [%]
RBC	-538.60	95.76	-634.35	99.2
MPC ($\lambda_{\text{wear}} = 0.8$)	-579.06	1.59	-580.65	100.0
MPC ($\lambda_{\text{wear}} = 0.4$)	-562.34	29.46	-591.80	99.7
MPC ($\lambda_{\text{wear}} = 0.3$)	-538.83	76.55	-615.38	99.3
MPC ($\lambda_{\text{wear}} = 0.03$)	-525.71	137.47	-663.18	98.8
RL ($w_{\text{wear}} = 0.6$)	-577.88	5.68	-583.56	100.0
RL ($w_{\text{wear}} = 0.5$)	-573.80	15.20	-589.00	99.9
RL ($w_{\text{wear}} = 0.4$)	-569.67	25.34	-595.01	99.8
RL ($w_{\text{wear}} = 0.2$)	-559.24	55.90	-615.15	99.5

controller well suited to real time operation and deployment on edge devices with limited computational resources.

From the analysis of the time evolution of the battery energy requests, reported in Fig. 5, it can be observed that the reinforcement learning controller produces actions with a smaller depth of discharge. This suggests that the learned PPO policy has more effectively internalized the higher penalty associated with deeper cycling, as encoded in the wear-cost model. Moreover, as shown in Fig. 6, a representative PPO training run exhibits a consistent learning trend over the course of the episodes, since the episode return increases, while the accumulated battery wear cost decreases. This indicates that the learned policy progressively improves both task performance and operational efficiency, converging toward a more effective control strategy that achieves better outcomes with lower battery degradation.

When focusing on the economic/wear tradeoff (Pareto perspective), MPC tends to achieve better solutions as the wear cost scaling is reduced, i.e., when the controller is progressively incentivized to exploit the battery more aggressively. In this regime, MPC can generate sharper charge/discharge actions and better time-shift energy thanks to oracle forecasts over the optimization horizon. Conversely, the RL policies form a less steep frontier. In fact, as one moves towards op-

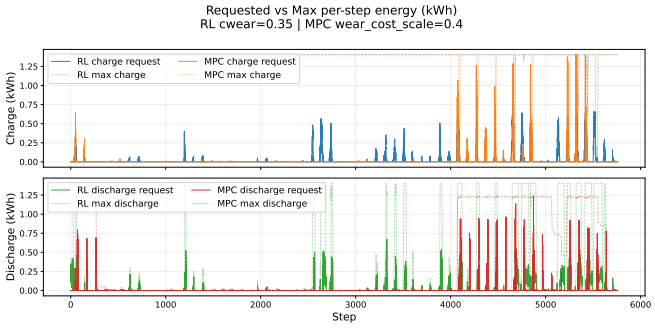

Fig. 5. Charge (top) and discharge (bottom) energy requests of the MPC and RL models over time.

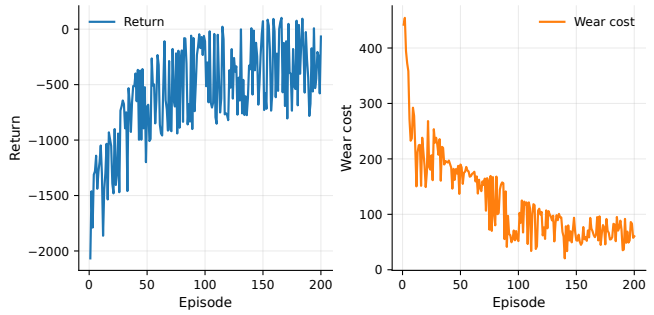

Fig. 6. Representative PPO training curves: episode return (left) and cumulative wear cost (right).

erating points with higher battery utilization (and thus higher wear), the gap with respect to MPC increases. An explanation of this result is given by the limited foresight of the RL agent in the considered setup. Indeed, without explicit forecasts, decisions are based on the current observation (including instantaneous prices and time features), which can lead to myopic behaviours such as importing from the grid while missing opportunities to store PV surplus or to discharge ahead of high-price periods. These differences become more pronounced in battery-intensive regimes, where anticipating near future conditions is critical for approaching oracle MPC performance.

IX. Conclusions

This work introduced a simulation framework for residential microgrid EMS with higher fidelity battery operation and aging than constant efficiency models. Aging is updated from experimentally identified SoH-Ah throughput curves, and a DoD based wear cost is included in the EMS objective. On top of this simulator, a PPO-based EMS learns a degradation-aware policy from interaction without load/PV forecasts. On a real household dataset, the best PPO configurations consistently outperform the rule-based baseline and approach oracle MPC at medium battery utilization; the gap increases in more aggressive regimes where lookahead becomes critical. Future work will incorporate load, PV, and price forecasts (including probabilistic forecasts) to model uncertainty and learn policies that are robust to forecast errors. Moreover, we aim to extend the assessment of the policy's economic impact over the entire battery lifecycle and explore multi-objective RL formulations to better balance conflicting operational goals.

X. Acknowledgments

This study was carried out within the MOST – Sustainable Mobility Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] Open Power System Data, "Household data," Available: https://data.open-power-system-data.org/household_data/2020-04-15/, 2020, version 2020-04-15.

[2] E. De Santis, A. Rizzi, and A. Sadeghian, "A learning intelligent system for classification and characterization of localized faults in smart grids," in *2017 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2017, pp. 2669–2676.

[3] M. Saleem, S. Saha, U. Izhar, and L. Ang, "A stochastic mpc-based energy management system for integrating solar pv, battery storage, and ev charging in residential complexes," *Energy and Buildings*, vol. 325, p. 114993, 2024.

[4] M. Amini, M. H. Nazari, and S. H. Hosseinian, "Optimal energy management of battery with high wind energy penetration: A comprehensive linear battery degradation cost model," *Sustainable Cities and Society*, vol. 93, p. 104492, 2023.

[5] O. Talab and İ. Avci, "Energy management in microgrids using model-free deep reinforcement learning approach," *IEEE Access*, vol. PP, pp. 1–1, 01 2025.

[6] O. N. Moga, A. Florea, C. Solea, and M. Vintan, "Reinforcement learning-based energy management in community microgrids: A comparative study," *Sustainability*, vol. 17, no. 23, 2025. [Online]. Available: https://www.mdpi.com/2071-1050/17/23/10696

[7] Y. Wei, S. Wang, X. Han, L. Lu, W. Li, F. Zhang, and M. Ouyang, "Toward more realistic microgrid optimization: Experiment and high-efficient model of li-ion battery degradation under dynamic conditions," *eTransportation*, vol. 14, p. 100200, 2022.

[8] G. Henri, T. Levent, A. Halev, R. Alami, and P. Cordier, "pymgrid: An open-source python microgrid simulator for applied artificial intelligence research," *CoRR*, vol. abs/2011.08004, 2020.

[9] M. A. I. Martins, L. B. Rhode, and A. B. D. Almeida, "A novel battery wear model for energy management in microgrids," *IEEE Access*, vol. 10, pp. 30 405–30 413, 2022.

[10] M. Ecker, N. Nieto, S. Käbitz, J. Schmalstieg, H. Blanke, A. Warnecke, and D. U. Sauer, "Calendar and cycle life study of Li(NiMnCo)O2-based 18650 lithium-ion batteries," *Journal of Power Sources*, vol. 248, pp. 839–851, Feb. 2014.

[11] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of machine learning research*, vol. 22, no. 268, pp. 1–8, 2021.

[12] A. V. Waghmare, V. P. Singh, T. Varshney, and P. Sanjeevikumar, "A systematic review of reinforcement learning-based control for microgrids: trends, challenges, and emerging algorithms," *Discover Applied Sciences*, vol. 7, no. 9, p. 939, 2025.

[13] M. Barbieri, M. Ceraolo, G. Lutzemberger, C. Scarpelli, T. Pesso, and M. Giovannucci, "Simplified electro-thermal model for lithium cells based on experimental tests," in *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, Turin, Italy, 2020, pp. 1–6.

[14] P. P. Mishra, A. Latif, M. Emmanuel, Y. Shi, K. McKenna, K. Smith, and A. Nagarajan, "Analysis of degradation in residential battery energy storage systems for rate-based use-cases," *Applied Energy*, vol. 264, p. 114632, 2020.