Deep Reinforcement Learning for Battery Arbitrage with Price and Renewable Data

Gabrielly de Q. Pereira[1], Marcella Scoczynski[2], Adriana A. Guimarães[2], Isabelly dos S. Pacheco[1], James W. Freitas[1], Diego Oliva[3], Felipe S. Silva[1]

Departamento de Informática - Universidade Estadual de Ponta Grossa (UEPG)
Ponta Grossa - PR - Brasil
Universidade Tecnológica Federal do Paraná (UTFPR) - Paraná - Brasil
Universidad de Guadalajara (UdeG) - Guadalajara - Jalisco - México

marcella@utfpr.edu.br, adriana@utfpr.edu.br, Isabellypacheco70@gmail.com, 45james46@gmail.com, diego.oliva@cucei.udg.mx

Abstract. *This paper proposes a Proximal Policy Optimization (PPO) approach for hourly battery energy arbitrage using open data from the Open Power System Data platform. The model uses day-ahead prices, load, renewable generation, and weather data from Germany and the Germany–Luxembourg price zone to learn continuous charging and discharging decisions under battery operational constraints. Using chronological training, validation, and test sets, PPO was compared with a price-quantile heuristic. The proposed policy achieved profits of 730.17 EUR in validation and 1019.15 EUR in testing, compared with 510.95 EUR and 491.78 EUR for the benchmark. The results demonstrate the potential of PPO for data-driven battery dispatch in day-ahead electricity markets.*

Resumo. *Este artigo propõe uma abordagem baseada em Proximal Policy Optimization (PPO) para arbitragem horária de energia com baterias utilizando dados abertos da plataforma Open Power System Data. O modelo utiliza preços do mercado do dia seguinte, carga, geração renovável e dados meteorológicos da Alemanha e da zona Alemanha–Luxemburgo para aprender decisões contínuas de carga e descarga sob restrições operacionais da bateria. Utilizando conjuntos cronológicos de treinamento, validação e teste, o PPO foi comparado com uma heurística baseada em quantis de preço. A política proposta alcançou lucros de 730,17 EUR na validação e 1.019,15 EUR no teste, frente a 510,95 EUR e 491,78 EUR da referência. Os resultados demonstram o potencial do PPO para o despacho de baterias orientado a dados em mercados de eletricidade do dia seguinte.*

1. Introduction

The operation of battery storage systems in energy markets has been studied as a sequential decision-making problem, in which the purchase and sale of energy depend on the temporal variation of prices and on the physical constraints of the asset [Harrold et al. 2022a, Cao et al. 2020, Miao et al. 2021]. The use of deep reinforcement learning in this context has been associated with the formulation of battery dispatch as a Markov decision process, allowing policies to be learned directly from historical time series without relying on an explicit market model [Cao et al. 2020, Zhang et al. 2024].

This line of research has also been applied in scenarios with renewable generation, variable demand, and dynamic prices, in which the battery acts as an element of operational flexibility [Harrold et al. 2022a, Jeong et al. 2023, Chen et al. 2024].

In arbitrage problems, part of the literature considers only the electricity price as the main input of the control policy, while other works incorporate forecasts, renewable generation, or meteorological information into the state space [Cao et al. 2020, Sage et al. 2024, Chen et al. 2024]. Harrold et al. applied Rainbow Deep Q-Networks to battery control in a microgrid with its own demand, renewable generation, and dynamic market prices, including scenarios with demand and price forecasts produced from local meteorological data [Harrold et al. 2022a]. Cao et al. formulated storage arbitrage with battery degradation in a United Kingdom wholesale market and combined a reinforcement learning method with price forecasting based on CNN and LSTM [Cao et al. 2020]. Jeong et al. jointly addressed the problem of renewable energy bidding and real-time battery control, considering the compensation of forecast errors through storage [Jeong et al. 2023]. Sage et al. analyzed arbitrage with deep reinforcement learning and time-series forecasting methods, presenting a case study with Alberta market prices and multiple prediction horizons [Sage et al. 2024]. Chen et al. evaluated the use of DRL in grid-connected photovoltaic-battery systems in buildings, including solar radiation in the state to examine the interaction between photovoltaic penetration, price volatility, and battery operation [Chen et al. 2024].

Another part of the literature broadens the scope of the problem to include additional mechanisms beyond pure arbitrage [Miao et al. 2021, Samende et al. 2022, Harrold et al. 2022b]. Miao et al. studied the co-optimization of energy arbitrage and frequency regulation in real-time markets, modeling the problem at multiple time scales [Miao et al. 2021]. Samende et al. considered a hybrid storage system with battery and hydrogen in an energy network with flexible demand, jointly addressing price, renewable production, and consumption [Samende et al. 2022]. Harrold et al. also investigated renewable integration and trading in microgrids with a multi-agent formulation, in which storage participates in an environment with cooperation and competition among agents [Harrold et al. 2022b].

There are also studies focused on incorporating additional operational constraints and specific forms of policy regularization [Xu et al. 2024, Jiang et al. 2025]. Xu et al. proposed an approach guided by offline optimal solutions for online storage control, using PPO and DDPG in short training horizons and adding noise to state transitions [Xu et al. 2024]. Jiang et al. proposed a dispatch method with protection against risk of loss in day-ahead markets, including action projection into feasible regions and behavior cloning from a teacher model [Jiang et al. 2025]. Zhang et al. studied energy management with battery degradation, load and price forecasting by GRU, and reinforcement learning with DQN in a smart grid with bilateral energy flow [Zhang et al. 2024].

Taken together, these studies show that the state formulation, the treatment of exogenous uncertainties, and the way battery constraints are represented vary according to the type of market, the service provided by storage, and the available data [Harrold et al. 2022a, Cao et al. 2020, Chen et al. 2024, Xu et al. 2024]. It is also observed that part of the literature uses microgrid environments, real-time markets, combined services, or dedicated forecasting models, while other approaches focus on on-

line control with observed historical time series [Harrold et al. 2022a, Miao et al. 2021, Sage et al. 2024, Xu et al. 2024].

In this work, the problem is treated as hourly battery arbitrage in the day-ahead market using open historical data from the Open Power System Data platform, with price, load, wind generation, solar generation, and meteorological variables aggregated for Germany and Germany-Luxembourg [Open Power System Data 2020a, Open Power System Data 2020b]. The proposed formulation uses PPO with a continuous action space, explicit representation of the state of charge, penalization for variation between consecutive actions, and a cost proportional to the battery energy throughput [Schulman et al. 2017]. Unlike the cited studies that address microgrids, service co-optimization, hybrid systems, forecasting through dedicated networks, or multi-agent structures, this work concentrates the analysis on a single battery arbitrage environment, OPSD data at hourly resolution, and price, renewable, load, and climate variables in the same training and evaluation pipeline [Harrold et al. 2022a, Cao et al. 2020, Jeong et al. 2023, Miao et al. 2021, Samende et al. 2022, Harrold et al. 2022b, Sage et al. 2024, Xu et al. 2024, Chen et al. 2024]. The objective of the article is to develop and evaluate a PPO-based dispatch policy for battery energy arbitrage, examine the effect of the set of price and renewable variables on sequential decision-making, and compare the performance of the learned policy with a heuristic benchmark based on price quantiles [Schulman et al. 2017, Open Power System Data 2020a, Open Power System Data 2020b].

2. Methodology

This work models the battery energy arbitrage problem as a sequential decision-making process in discrete time, in which an agent chooses, at each hour, how much to charge or discharge based on historical information on price, load, renewable generation, and meteorological variables. The agent's objective is to maximize the accumulated economic return over the operation horizon, while respecting physical constraints of battery capacity, power, and efficiency. The computational implementation was developed in Python using the libraries `gymnasium`, `stable-baselines3`, `PyTorch`, `pandas`, and `scikit-learn`. The reinforcement learning algorithm adopted was Proximal Policy Optimization (PPO), following the formulation proposed by [Schulman et al. 2017].

The data used were obtained from the Open Power System Data (OPSD) platform [Open Power System Data 2020c, Open Power System Data 2020a, Open Power System Data 2020b], which provides datasets for power system modeling with standardized documentation, versioning, and formats suitable for reproducible experiments. In this study, two OPSD packages were combined, including the *Time series* package and meteorological variables aggregated by country at hourly resolution. The selected variables correspond to Germany and the Germany–Luxembourg price zone and included DE_load_actual_entsoe_transparency, DE_load_forecast_entsoe_transparency, DE_solar_generation_actual, DE_wind_generation_actual, DE_LU_price_day_ahead, DE_temperature, DE_radiation_direct_horizontal, and DE_radiation_diffuse_horizontal. The merge between the two packages was performed using the utc_timestamp field through an *inner join*, preserving only timestamps common to both datasets.

After integrating the datasets, the records were chronologically ordered and restricted to the period between 2018-10-01 and 2019-12-31, corresponding to the selected common time window for the experiment. Numerical variables with missing values were handled through linear interpolation over time, with bidirectional filling, and any remaining rows with missing data were removed. This step aimed to produce a continuous time series for the simulation of the decision environment.

The system state representation was built from observed and derived variables. Let p_t denote the day-ahead price at time t, L_t the observed load, $\hat{L}_t$ the day-ahead load forecast, S_t the solar generation, W_t the wind generation, T_t the temperature, R_t^d the direct horizontal radiation, and R_t^f the diffuse horizontal radiation. From these quantities, auxiliary variables were constructed to enrich the operational context. The aggregated renewable generation was defined as

$$G_t^{\text{ren}} = S_t + W_t, \tag{1}$$

while the net load was defined as

$$L_t^{\text{net}} = L_t - G_t^{\text{ren}}. \tag{2}$$

The share of renewable generation in the load was also computed as

$$\phi_t^{\text{ren}} = \frac{G_t^{\text{ren}}}{L_t + 10^{-6}}, \tag{3}$$

as well as the load forecast error,

$$e_t^L = \hat{L}_t - L_t. \tag{4}$$

To incorporate aggregated climatic information, total radiation was defined as

$$R_t^{\text{tot}} = R_t^d + R_t^f. \tag{5}$$

Additionally, temporal price differences over one hour and twenty-four hours were constructed: $\Delta p_t^{(1)} = p_t - p_{t-1}$,
$\Delta p_t^{(24)} = p_t - p_{t-24}$.

Temporal lags of price, load, renewable generation, temperature, and total radiation were also included, with delays of 1, 2, 3, 6, 12, and 24 hours. The goal was to provide the agent with information on short-term temporal dependencies, intraday variations, and recurring daily patterns. Since price, load, and renewable generation exhibit hourly, weekly, and monthly seasonality, time was encoded using cyclic variables. Thus, hour of day, day of week, and month were transformed into sine and cosine pairs: $\text{h}_t^{\sin} = \sin\left(2\pi \frac{\text{hour}(t)}{24}\right), h_t^{\cos} = \cos\left(2\pi \frac{\text{hour}(t)}{24}\right)$,
$d_t^{\sin} = \sin\left(2\pi \frac{\text{dayOfWeek}(t)}{7}\right), d_t^{\cos} = \cos\left(2\pi \frac{\text{dayOfWeek}(t)}{7}\right)$,
$m_t^{\sin} = \sin\left(2\pi \frac{\text{month}(t)}{12}\right), m_t^{\cos} = \cos\left(2\pi \frac{\text{month}(t)}{12}\right)$. A binary weekend indicator was also included. Thus, the exogenous feature vector represents not only instantaneous values but also temporal positioning and recent history.

The dataset was split chronologically. The training period included observations up to 2019-06-30, the validation period covered 2019-07-01 to 2019-09-30, and the test

period covered 2019-10-01 to 2019-12-31. This split preserves temporal order and avoids leakage of future information into training. After feature construction, standardization to zero mean and unit variance was applied using `StandardScaler`. The scaler was fitted only on the training set and then applied to validation and test sets, preventing future information leakage.

The reinforcement learning environment was implemented as a custom class named `BatteryArbitrageEnv`, compatible with the `gymnasium` interface. The battery was parameterized with maximum energy capacity $E^{\max} = 1.0$ MWh, power limit $P^{\max} = 0.25$ MW, charge efficiency $\eta_c = 0.96$, and discharge efficiency $\eta_d = 0.96$. At the start of each episode, the state of charge was fixed at 0.5 MWh (50% capacity). The observation vector consisted of 39 components: 36 exogenous/derived features plus three operational variables (state of charge, previous action, and episode progress). Thus, the state at time t is

$$s_t = [x_t, \mathrm{SoC}_t, a_{t-1}, \rho_t], \tag{6}$$

where $x_t \in R^{36}$, SoC_t is the state of charge, a_{t-1} the previous action, and $\rho_t \in [0, 1]$ the relative episode progress.

The action variable is continuous:

$$a_t \in [-1, 1]. \tag{7}$$

It is mapped to power as

$$P_t = a_t P^{\max}. \tag{8}$$

The reward function is

$$r_t = e_t^d p_t - e_t^c p_t - c_{\deg}(e_t^c + e_t^d) - c_{\mathrm{sm}}|a_t - a_{t-1}|. \tag{9}$$

The PPO algorithm uses the clipped objective:

$$L^{\mathrm{CLIP}}(\theta) = \hat{E}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right], \tag{10}$$

The Sharpe-like metric is:

$$\mathrm{SR} = \frac{\bar{r}}{\sigma_r + 10^{-8}} \sqrt{24 \times 365}, \tag{11}$$

The heuristic policy is:

$$a_t = \{\, 1\ , p_t \leq Q_{0.30},\ -1, p_t \geq Q_{0.70}, 0, otherwise. \tag{12}$$

[H] [1] Load OPSD *Time series* and *Weather Data* Select load, forecast, wind, solar, price, and climate variables Merge datasets using UTC timestamp Interpolate missing values and remove incomplete rows Build derived variables, lags, and cyclic encodings Split data into train, validation, and test sets chronologically Fit `StandardScaler` on training data and transform others Initialize `BatteryArbitrageEnv` Apply `VecNormalize` Initialize PPO with MLP policy $k = 1$ to training iterations Collect rollouts of 2048 steps Compute advantages with GAE Optimize PPO objective evaluation step reached Evaluate on validation set Save best model Apply early stopping if needed Reload best model Generate trajectories for validation and test Compute performance metrics Evaluate heuristic policy Export tables and plots

Tabela 1. Configuration of the PPO model and arbitrage environment

Item	Value
Algorithm	PPO
Policy	`MlpPolicy`
Network architecture	2 hidden layers, 256 neurons each
Activation function	ReLU
Learning rate	3×10^{-4}
Rollout length	2048 steps
Batch size	256
Epochs per rollout	10
Discount factor γ	0.995
GAE λ	0.98
Clip range	0.2
Entropy coefficient	0.005
Value function coefficient	0.5
Gradient clipping	0.5
Random seed	42
Training budget	120000 *timesteps*
Executed steps	120832
Battery capacity	1.0 MWh
Maximum power	0.25 MW
Charge efficiency	0.96
Discharge efficiency	0.96
Degradation cost	2.0 EUR/MWh
Smoothing penalty	0.05

3. Results and Discussion

This section reports the performance of the PPO-based dispatch policy for hourly battery arbitrage and compares it with the quantile-based heuristic benchmark. The analysis is based on the exported artifacts of the experiment, including aggregated performance tables, step-by-step dispatch trajectories, and graphical summaries of learning dynamics, cumulative profit, and battery operation over time.

Table 1 summarizes the experimental configuration adopted in the study. The control policy was implemented as an MLP with two hidden layers of 256 neurons and `ReLU` activation. Training was configured with a budget of 120000 *timesteps*, and the final execution reached 120832 environment interaction steps because PPO updates are performed in rollout blocks of 2048 steps. The battery model used in the environment had a capacity of 1.0 MWh, a power limit of 0.25 MW, charge and discharge efficiencies of 0.96, a degradation cost of 2.0 EUR/MWh, and a smoothing penalty of 0.05 on action variation between consecutive steps.

Table 2 presents the main characteristics of the chronological splits used in the experiment. The training set contains 6528 hourly observations, and the validation and test sets contain 2208 observations each. The training interval exhibits the highest average price level, while the validation set corresponds to a warmer seasonal regime and the test set corresponds to a colder regime. These differences indicate that the learned policy was

Tabela 2. Statistical summary of the train, validation, and test sets

Set	Hours	Mean price	Std. dev.	Mean load	Mean renewable
Train	6528	43.07	19.18	57389.18	19484.02
Validation	2208	37.45	11.63	53293.68	16483.48
Test	2208	36.59	14.36	56595.61	18693.06

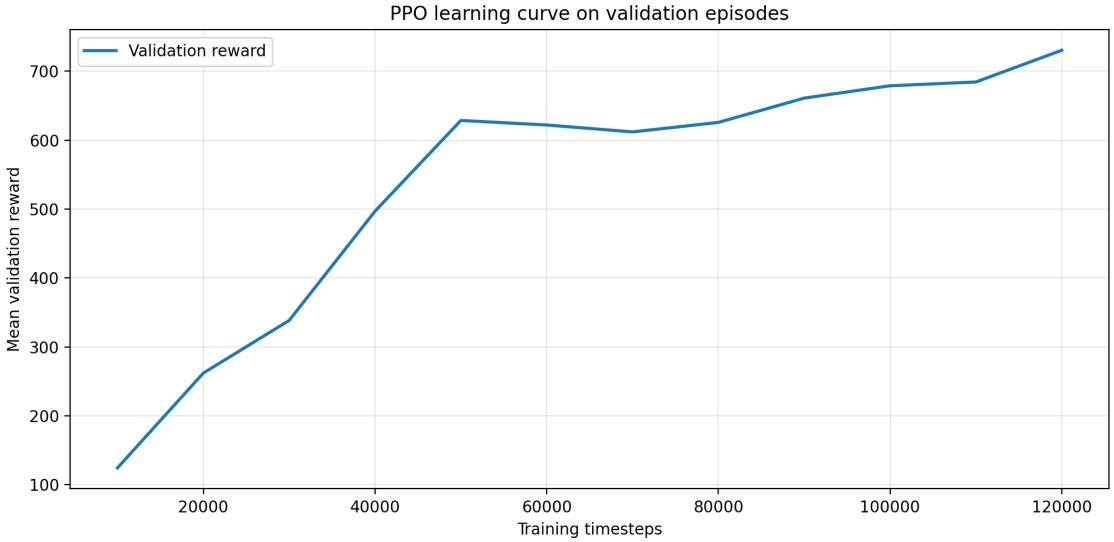


Figura 1. PPO learning curve on the validation set.

evaluated under temporal conditions that differ from those observed during training.

The learning dynamics are shown in Fig. 1. The validation reward increased from 124.43 at 10000 training steps to 730.47 at 120000 steps. Between 50000 and 80000 steps, the curve entered a region with smaller variation, ranging approximately from 612.03 to 628.62, and then resumed increasing in the later stage of training. This pattern indicates that the policy continued improving after the initial training phase and that the later updates contributed to additional gains in out-of-sample validation return.

The main economic and operational results are reported in Tables 3 and 4. In the validation interval, PPO achieved a total profit of 730.17 EUR, whereas the quantile baseline obtained 510.95 EUR. In the test interval, PPO obtained 1019.15 EUR, while the baseline obtained 491.78 EUR. The corresponding mean hourly reward was 0.3307 EUR/h and 0.4616 EUR/h for PPO in validation and test, compared with 0.2314 EUR/h and 0.2227 EUR/h for the baseline. The Sharpe-like ratio was also higher for PPO in both intervals, reaching 8.24 in the test period compared with 6.20 for the benchmark.

To make the difference more explicit, Table 5 reports the relative improvement of PPO over the benchmark. The learned policy increased total profit by 42.90% in validation and 107.24% in test. The same relative gains appear in the mean hourly reward because both metrics are directly proportional to the accumulated reward over the fixed evaluation horizon. The Sharpe-like ratio also improved by 11.18% in validation and 32.93% in test.

The results in Tables 3–5 show that the advantage of PPO is not restricted to

Tabela 3. Economic performance comparison

Model	Split	Profit (EUR)	Mean r (EUR/h)	Std r (EUR/h)	SR
PPO	Validation	730.17	0.331	5.008	6.181
PPO	Test	1019.15	0.462	5.242	8.242
Quantile	Validation	510.95	0.231	3.896	5.559
Quantile	Test	491.78	0.223	3.362	6.200

Tabela 4. Operational performance comparison

Model	Split	Charged (MWh)	Discharged (MWh)	Final SoC (MWh)	Steps
PPO	Validation	90.49	83.37	0.524	2208
PPO	Test	103.52	95.39	0.507	2208
Quantile	Validation	45.96	42.61	0.240	2208
Quantile	Test	35.80	33.26	0.219	2208

Tabela 5. Relative improvement of PPO over the quantile baseline

Split	Profit gain (%)	Mean reward gain (%)	SR gain (%)
Validation	42.90	42.90	11.18
Test	107.24	107.24	32.93

Tabela 6. Operational behavior of PPO and quantile baseline

Model	Split	$\|a\|$ mean	Charge (%)	Discharge (%)	Idle (%)
PPO	Validation	0.759	29.39	70.61	0.00
PPO	Test	0.798	41.58	58.42	0.00
Quantile	Validation	0.601	30.07	30.03	39.90
Quantile	Test	0.601	30.03	30.03	39.95

accumulated profit. The learned controller also made more extensive use of the battery. In the test set, PPO charged 103.52 MWh and discharged 95.39 MWh over 2208 hours, while the heuristic policy charged 35.80 MWh and discharged 33.26 MWh. At the same time, the final state of charge of PPO remained close to the initial operating level of 0.5 MWh, indicating that the higher profit was not obtained by ending the horizon in an artificially favorable energy state.

To complement the aggregated performance metrics, Table 6 reports the action distribution of the two policies. PPO produced nonzero actions throughout the entire evaluation horizon, whereas the quantile policy remained inactive for approximately 40% of the hours in both validation and test. In the test set, PPO issued positive actions in 41.58% of the hours and negative actions in 58.42% of the hours, with a mean absolute action value of 0.798. These values indicate that the learned controller used the continuous action space more actively than the heuristic rule.

The state-of-charge statistics reported in Table 7 further clarify how the battery was used. PPO employed the full operational interval of the battery, reaching both 0 and 1.0 MWh at some points in the horizon. In the test set, the mean SoC of PPO was 0.385

Tabela 7. State-of-charge statistics

Model	Split	Mean SoC	Min SoC	Max SoC	Final SoC
PPO	Validation	0.212	0.000	1.000	0.524
PPO	Test	0.385	0.000	1.000	0.507
Quantile	Validation	0.453	0.000	1.000	0.240
Quantile	Test	0.479	0.000	1.000	0.219

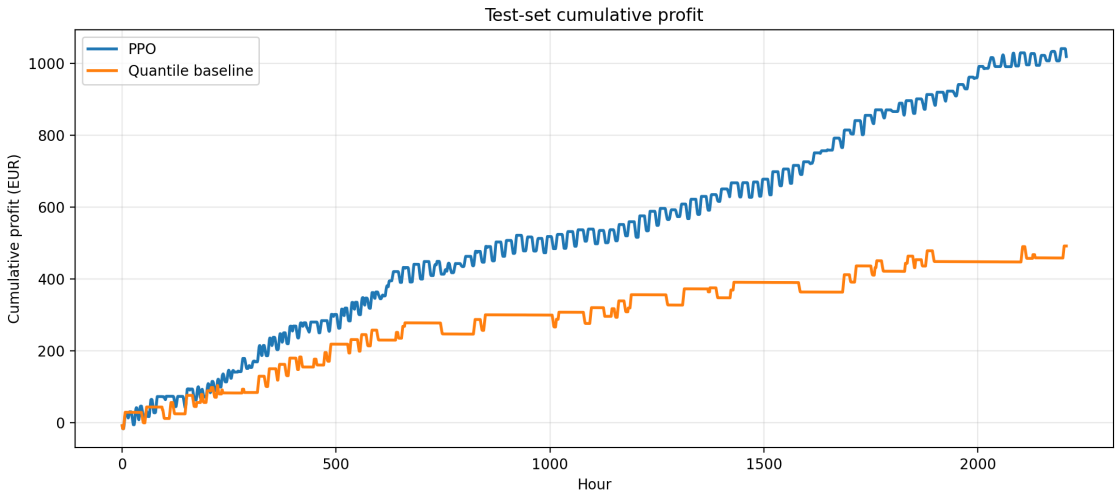


Figura 2. Cumulative profit on the test set for PPO and the quantile baseline.

MWh, whereas the quantile baseline remained at a higher mean SoC of 0.479 MWh and with lower total throughput. This difference suggests that PPO learned a more dynamic use of storage over time, rather than remaining concentrated around a narrower operating pattern.

Fig. 2 shows the cumulative profit in the test interval. The PPO trajectory remains above the baseline trajectory at the end of the period, and the gap between the two policies increases over time. This behavior indicates that the learned policy identified charge and discharge patterns that the fixed threshold rule did not capture. The varying slope of the PPO curve is consistent with a dispatch policy that responds to hourly price conditions together with battery state and exogenous context variables.

The step-by-step trajectories also help clarify what the trained model predicts at inference time. The PPO agent does not directly predict future prices or renewable generation. Instead, at each hour, it maps the observed state vector into a continuous action $a_t \in [-1, 1]$, which represents the battery charging or discharging command. The quality of this predicted action is evaluated indirectly through its operational consequences, namely charged energy, discharged energy, updated state of charge, hourly reward, and cumulative profit. In the exported results, these quantities were stored for every step of the validation and test intervals.

The input to the learned policy includes day-ahead price, actual load, load forecast, solar generation, wind generation, temperature, direct radiation, diffuse radiation, engineered variables such as renewable share and net load, temporal lags, calendar encodings,

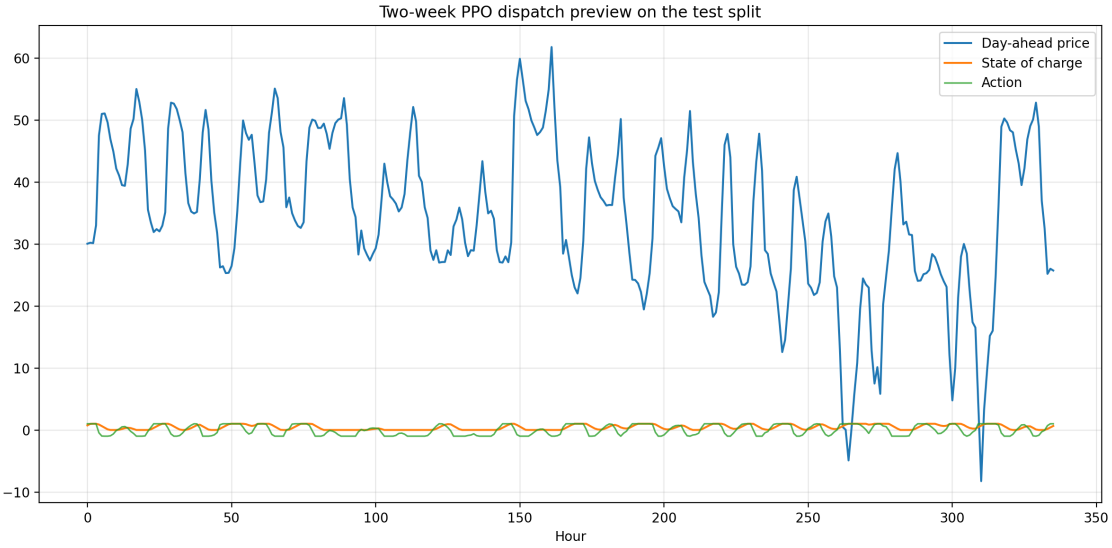


Figura 3. Two-week window of the test set with day-ahead price, agent action, and battery state of charge.

the current state of charge, the previous action, and the episode progress. Therefore, the prediction generated by the model at each hour reflects a joint use of market, renewable, climatic, and internal storage information. The improved results of PPO relative to the baseline indicate that this richer state representation contributed to the dispatch decisions learned by the policy.

Fig. 3 presents a two-week excerpt of the test interval with day-ahead price, action, and state of charge. The figure shows the temporal relation between price variation and the battery response. When the controller selects charging actions, the state of charge increases until the battery reaches its physical limit; when discharging actions are selected, the stored energy decreases consistently with the power bound and the available energy in the battery. This figure illustrates that the learned policy follows the physical constraints embedded in the environment while adapting dispatch intensity over time.

From the perspective of reward dispersion, PPO presented a higher hourly standard deviation than the baseline. In the test set, the standard deviation of hourly reward was 5.24 EUR for PPO and 3.36 EUR for the heuristic rule. This result needs to be interpreted jointly with the higher mean return obtained by PPO. Since the mean reward and total profit were both higher, and the Sharpe-like ratio also increased, the larger dispersion is associated with a policy that exploits more trading opportunities over time rather than a deterioration in the economic profile measured by the adopted summary metrics.

Another point of interest is the relation between validation and test performance. PPO obtained 730.17 EUR in validation and 1019.15 EUR in test, while the baseline obtained 510.95 EUR and 491.78 EUR, respectively. The learned policy therefore preserved its advantage in both intervals and increased the difference relative to the benchmark in the out-of-sample test period. Under the chronological split adopted in this study, this result is consistent with generalization across different market and seasonal conditions within the historical horizon considered.

With respect to application, the proposed dispatch policy can be used in studies of grid-connected battery operation in day-ahead settings with hourly prices. Examples include storage operated by aggregators, traders, microgrids, industrial consumers with battery assets, and renewable plants coupled with storage. In systems with wind or photovoltaic generation, the same control logic can be used to decide when to absorb renewable surplus and when to shift energy delivery toward higher-price periods. In centralized operational studies, the framework can also support scenario comparison across battery sizes, power limits, efficiencies, and market conditions using the same learning pipeline.

The environment analyzed here was restricted to a single battery and a single day-ahead market with linear degradation cost and fixed efficiency. Within these assumptions, the comparison with the quantile benchmark shows that PPO extracted more return from the same observation set under the same physical constraints. The main distinction between the two approaches lies in the temporal structure of the actions. The quantile policy only reacts when price crosses fixed thresholds, whereas PPO continuously adjusts the action as a function of the full observed state, including price, renewables, load, weather, temporal memory, and internal battery state.

Overall, the results show that the PPO agent learned a dispatch policy with higher total profit, higher mean hourly reward, and higher Sharpe-like ratio than the heuristic benchmark, in both validation and test intervals. The exported trajectories and graphical summaries also show that the learned policy used the battery more actively, respected the state and power constraints, and maintained a feasible state-of-charge profile throughout the horizon. These results support the use of the proposed framework for data-driven studies of battery arbitrage with open hourly price, renewable, load, and weather data.

4. Conclusion

This paper presented a reinforcement learning framework for hourly battery energy arbitrage using open historical data from the Open Power System Data platform. The proposed approach integrates day-ahead prices, load, renewable generation, and weather variables within a continuous-action PPO policy, while considering battery capacity, power limits, efficiencies, degradation costs, and action smoothing. The evaluation used chronological training, validation, and test splits and a heuristic benchmark based on price quantiles.

The PPO policy outperformed the heuristic benchmark in both evaluation periods, achieving profits of 730.17 EUR versus 510.95 EUR in validation and 1019.15 EUR versus 491.78 EUR in testing. It also achieved higher mean hourly reward and Sharpe-like ratio while maintaining feasible state-of-charge behavior. These results demonstrate the potential of PPO to learn effective battery dispatch strategies directly from historical market and system data.

Future work may investigate other price zones, battery configurations, forecast-based state representations, multi-service operation, and more detailed degradation models. Comparisons with other continuous-control reinforcement learning algorithms may further assess the robustness and generalizability of the proposed approach.

Referências

Cao, J., Harrold, D., Fan, Z., Morstyn, T., Healey, D., and Li, K. (2020). Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model. *IEEE Transactions on Smart Grid*, 11(5):4513–4521.

Chen, Q., Kuang, Z., Liu, X., and Zhang, T. (2024). Application-oriented assessment of grid-connected pv-battery system with deep reinforcement learning in buildings considering electricity price dynamics. *Applied Energy*, 364:123163.

Harrold, D. J. B., Cao, J., and Fan, Z. (2022a). Data-driven battery operation for energy arbitrage using rainbow deep reinforcement learning. *Energy*, 238:121958.

Harrold, D. J. B., Cao, J., and Fan, Z. (2022b). Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning. *Applied Energy*, 318:119151.

Jeong, J., Kim, S. W., and Kim, H. (2023). Deep reinforcement learning based real-time renewable energy bidding with battery control. *IEEE Transactions on Energy Markets, Policy and Regulation*, 1(2):85–96.

Jiang, X., Tang, F., Chen, M., Li, B., Yu, Y., Ding, J., and Li, X. (2025). A reinforcement learning-based approach with downside-risk protection for battery dispatch in day-ahead markets. *International Transactions on Electrical Energy Systems*.

Miao, Y., Chen, T., Bu, S., Liang, H., and Han, Z. (2021). Co-optimizing battery storage for energy arbitrage and frequency regulation in real-time markets using deep reinforcement learning. *Energies*, 14(24):8365.

Open Power System Data (2020a). Data package time series. `https://doi.org/10.25832/time_series/2020-10-06`. Version 2020-10-06.

Open Power System Data (2020b). Data package weather data. `https://doi.org/10.25832/weather_data/2020-09-16`. Version 2020-09-16.

Open Power System Data (2020c). Data platform. Accessed: 2026-05-05.

Sage, M., Campbell, J., and Zhao, Y. F. (2024). Enhancing battery storage energy arbitrage with deep reinforcement learning and time-series forecasting. In *Proceedings of the ASME 2024 Energy Sustainability Conference*, page V001T01A007.

Samende, C., Fan, Z., and Cao, J. (2022). Battery and hydrogen energy storage control in a smart energy network with flexible energy demand using deep reinforcement learning. *arXiv preprint arXiv:2208.12779*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Xu, G., Shi, J., Wu, J., Lu, C., Wu, C., Wang, D., and Han, Z. (2024). An optimal solutions-guided deep reinforcement learning approach for online energy storage control. *Applied Energy*, 361:122915.

Zhang, H., Zhang, G., Zhao, M., and Liu, Y. (2024). Load forecasting-based learning system for energy management with battery degradation estimation: A deep reinforcement learning approach. *IEEE Transactions on Consumer Electronics*, 70(1):2342–2352.