Subramanya, Rakshith; Aaltonen, Harri; Sierla, Seppo; Vyatkin, Valeriy

Onsite Renewable Generation Time Shifting for Photovoltaic Systems

Published in:
2023 IEEE 32nd International Symposium on Industrial Electronics, ISIE 2023 - Proceedings

DOI:
10.1109/ISIE51358.2023.10228097

Published: 01/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Subramanya, R., Aaltonen, H., Sierla, S., & Vyatkin, V. (2023). Onsite Renewable Generation Time Shifting for Photovoltaic Systems. In *2023 IEEE 32nd International Symposium on Industrial Electronics, ISIE 2023 - Proceedings* (Proceedings of the IEEE International Symposium on Industrial Electronics; Vol. 2023-June). IEEE. https://doi.org/10.1109/ISIE51358.2023.10228097

Onsite Renewable Generation Time Shifting for Photovoltaic Systems

Rakshith Subramanya
Department of Electrical Engineering and Automation
Aalto University
Espoo, Finland
rakshith.subramanya@aalto.fi

Aaltonen Harri
Department of Electrical Engineering and Automatin
Aalto University
Espoo, Finlad
harri.aaltonen@aalto.fi

Seppo Sierla
Department of Electrical Engineering and Automation
Aalto University
Espoo, Finland
seppo.sierla@aalto.fi

Valeriy Vyatkin
Department of Electrical Engineering and Automation, Department of Computer Science, Electrical and Space Engineering
Aalto University, Luleå University of Technology
Espoo, Finland. Luleå, Sweden
valeriy.vyatkin@aalto.fi

***Abstract*— This paper examines the challenge of bidding a battery system on an electrical power market with varying prices. Optimizing bids for each market interval is a complex challenge since the bid size for one market interval affects the battery availability during several subsequent market intervals. In this paper, a renewable generation unit such as Photovoltaic (PV) and a battery storage are considered, with the battery used to shift electricity sales from low-price market intervals to high-price market intervals. The battery can also be used to smooth momentary fluctuations in generated power, cope with differences between actual and forecasted generation, and help to meet the maximum power limit constraint. This paper evaluates how to best manage the battery for optimal bidding in an electrical power market.**

Keywords—battery, reinforcement learning, artificial intelligence, electricity market, frequency reserve.

I. Introduction

The market for electrical power is highly volatile, and prices frequently change. The supply and demand, weather, fuel costs, and other variables all affect price fluctuations in the electrical power market. In order to guarantee that consumers have access to dependable and reasonably priced electricity, utility companies are continuously looking for ways to keep prices low. This paper considers a renewable generation unit such as Photovoltaic (PV) generation and a battery, selling electricity to an electrical power market with prices that may vary from one market interval to the next. A battery may be used to shift the electricity sales to market intervals with higher prices [1], [2] .

With resources such as wind and photovoltaic generation, there can be momentary fluctuations in the generated power, and many electrical power markets accept such fluctuations as long as the total delivered energy during the market interval matches the amount specified in a bid made by the site operator to the market [3], [4]. However, there is a maximum power limit constraint at the site's grid coupling, which must be considered as a constraint for managing the power flows. The battery may be used to smooth the fluctuations in order to respect this constraint [5]. Thus, the problem of bidding on the market is concerned with the energy (kWh) capacity delivered during the market interval rather than real-time power (kW) fluctuations within the interval.

Since renewable energy sales are made using possibly erroneous generation forecasts, the battery can also be used to cope with differences between actual and forecasted generation [6]. If the battery becomes full or empty during this process, it may be impossible to deliver the sold energy to the grid without violating the power limit constraint, so the system may incur a penalty from the market [7], [8]. Further, the battery may be used to shift sales of electricity from low price market intervals to high price market intervals. On low price market intervals, the sell bid kWh capacity is set to be below the renewable generation forecast, so part of the generation is used to charge the battery. On high price market intervals, the bid capacity can be set higher than the generation forecast, so the battery will be discharged. Such bidding is complicated by the fact that in general, the price is not known at bidding time, so only an uncertain price estimate is available. Further still, any use of the battery will involve a cost due to battery aging, which depends on the battery State of Charge (SoC) and the charging and discharging currents [9].

So, relying on uncertain generation forecasts and market price forecasts, the problem of optimizing the bids for each market interval is not straightforward if all of the above-mentioned considerations are taken into account. It is notable that the bid size for one market interval can impact the use of the battery during that market interval, so it will impact the state of charge of the battery at the beginning of the next market interval. This in turn can impact the battery's ability to smooth fluctuations, cope with erroneous generation forecasts or to store energy to be sold later during a higher price market interval. Thus, it is not possible to optimize bids one market interval at a time, without considering the impact of the bid on system performance during several subsequent market intervals.

II. Literature Review

Agent, environment, state, action, and reward are the five fundamental components of Reinforcement learning (RL). The agent performs actions on the environment and updates model parameters according to its own state. For the performed actions, agent receives rewards from the environment [10]. RL can be utilized to carry out a multi-objective optimization in a system made up of renewable generation, battery storage, grid coupling, and a bidding optimizer with access to an electrical power market. The optimization could be targeted to achieve maximum market revenue, lowest possible market penalties and minimum battery aging. Dong et al. [11] use RL agent to trade the battery for an energy market by minimizing the battery aging costs and charging/discharging losses. To take advantage of distributed energy resources in multiple energy markets, De Silva et al. [12] suggest a machine learning architecture for a Virtual Power Plant (VPP) setup. According to Subramanya et al. [13], a VPP is in charge of such a commercially active installation that takes part in Finnish frequency reserve markets.

The machine learning based time series forecasts are easily accessible for markets like frequency reserve markets [14], real-time markets [15], and day-ahead spot markets [16]. Several authors have proposed numerous solutions to the energy arbitrage problem with battery and renewable generation. Bai et al. [17] optimize the Battery Energy Storage System (BESS) operations using the battery model while taking revenues and battery loss into account. In order to increase the income from arbitrage services, Garca-Miguel et al. [18] evaluate how the degradation cost of use should be taken into account.

This paper proposes a multi-objective optimization method for bidding and managing the power flows between the generation, storage and grid, which maximizes market revenue, and minimizes market penalties by –

- Using the battery storage to manage errors in generation forecasts used at bidding time.
- Ensuring that the maximum power flow limits at the grid coupling are never exceeded, despite fluctuations in renewable generation.
- Using the storage to shift sales of the generated renewable energy to market intervals with higher prices, even if only uncertain price forecasts are available at the time of placing the bids.

The designed system's architecture and the method for RL agent training are covered in Part III. The system description, RL environment generation, and agent interaction with the environment are all explained in Section IV. Results and analysis of the designed system are presented in Section V. The paper is concluded in Section VI.

III. ARCHITECTURE

A system for using the storage to shift sales of the generated renewable energy when uncertain market price forecasts are as is shown in Figure 1. In this paper, consideration is given specifically to wind and solar power.

The Electricity market is any electrical power market in which participants may place bids for selling electric energy, such as a day-ahead, intraday, or local market. The wind turbine or solar panel consists of one or more renewable generation units, such as wind or PV generation, producing Direct Current (DC) power. One or more grid tie inverters perform the DC/AC (Alternating Current) conversion and feeds the power to the grid. The controller of the wind park or solar plant receives measurement such as frequency and voltage from the grid and the grid tie inverter and controls the inverter output so that the grid code requirements of the Utility grid are met. Additionally, the controller is able to receive setpoints from a Bidding optimizer, and based on these setpoints, it can adjust DC power flows to the Battery storage or AC power flows to the Utility grid through the inverter and Grid coupling. The controller is also able to curtail the generation if needed. The Battery storage is able to charge the battery using the power received from the Renewable generation unit, and it is able to discharge to the Utility grid through the inverter and the Grid coupling. The Battery storage is controlled by a Battery Management System that accepts a control signal that specifies the setpoints for the power flows used to charge or discharge the battery.

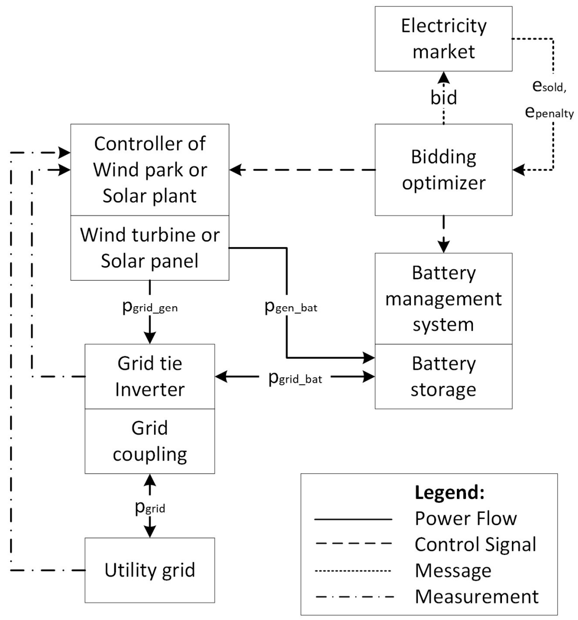

Fig. 1. System Architecture

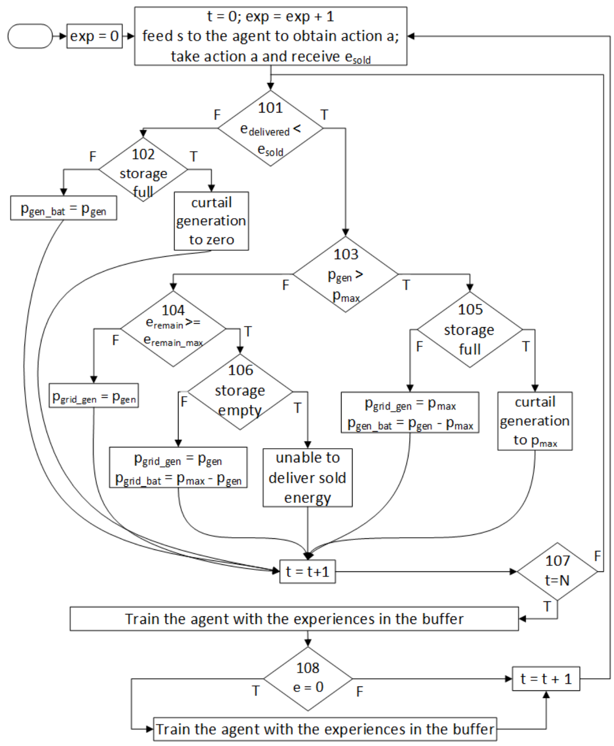

Fig. 2. Method for training a reinforcement learning agent to optimize the bidding and manage the power flows

The Battery storage may include a stationary battery or a vehicle to grid (V2G) battery. The Bidding optimizer bids on the Electricity market and sends the control signals to the Controller and Battery Management System.

The novelty of this paper is in the design of the Bidding optimizer, which enables a new way of coordinating with various systems such as Renewable generation, Grid inverters, Renewable generation controllers, Battery storage and Battery storage management systems. The novel method that is executed by Bidding optimizer is disclosed in Figure 2.

The bids are made for each market interval, and the bid consists of an energy in kWh which the participant commits to sell to the market as well as a price at which the participant is willing to sell it. Once a bid is accepted, the energy quantity in the bid is denoted as e_{sold}. An exemplary market interval is one hour, and in general terms, the duration of the market interval is T.

The market interval T is divided to N timesteps at which setpoint changes may be made by the Controller and the Battery Management System. The duration of one timestep is T/N. The current timestep is denoted with t. If the sold energy is delivered to the utility grid through the Grid coupling at a constant power p_{grid}, the following two equations hold:

$$p_{grid} = e_{sold} / T \tag{1}$$

$$p_{grid} = p_{grid\_bat} + p_{grid\_gen} \tag{2}$$

p_{gen} is used to denote the momentary generation of power by the Renewable generation. This can be directed to the grid or the battery or both:

$$p_{gen} = p_{grid\_gen} + p_{gen\_bat} \tag{3}$$

During a market interval, at timestep t, the energy that has been so far delivered to the grid is:

$$e_{delivered} = \sum_{i=0}^{t}(p_{grid} * \frac{T}{N}) \tag{4}$$

The energy that remains to be delivered during the market interval is:

$$e_{remain} = e_{sold} - e_{delivered} \tag{5}$$

The Grid coupling has a maximum power flow p_{max}, which is specified in the contract with the utility. Exceeding this maximum power, even for a short time, can involve a large cost, and thus must be avoided. At this maximum power, the energy that could be delivered to the grid during the remainder of the market interval is:

$$e_{remain\_max} = \sum_{i=t}^{T-1}(p_{max} * \frac{T}{N}) \tag{6}$$

The structure of the bid depends on the specific requirements of the Electricity market. In one implementation of this concept, the bid consists of an amount of energy e_{bid} to be sold to the grid during a market interval (e.g., in kWh) and the minimum price $price_{bid}$ at which the seller is willing to sell this energy. The bid is sent by the Bidding optimizer to the Electricity market, which returns the outcome of the bid, namely e_{sold} and the price at which the bid was accepted $price_{accept}$, which may be higher than $price_{bid}$, depending on what mechanism, such as an auction, is used by Electricity market to determine the price based on all of the received bids.

A reinforcement learning agent is trained to form the bids within Bidding optimizer, and a method for this purpose is disclosed in Figure 2. The input to the agent is a state data structure s. The s includes the generation forecast and market price forecast for the upcoming market interval and relative size of the forecasts within following 24 hours. The ratio of relative predicted prices (RPP) is calculated as:

$$RPP(t) = PP(t)/max(PP(t):PP(t+23)) \tag{7}$$

The ratio of relative PV production prediction (RPVPP) is calculated as:

$$RPVPP = PVPP (t)/max(PVPP (t): PVPP (t+23)) \tag{8}$$

Where PP is price prediction for the hour and $PVPP$ is the prediction of PV production for the hour. The s also includes normalized history data of battery usage from previous 10 hours. The output of the agent is an action a, which specifies the values for e_{bid} and $price_{bid}$. After taking the action a, the Bidding optimizer will receive e_{sold} and $price_{accept}$. Based on e_{sold}, it will form the control signals that specify the setpoints of the power flows in Figure 1. The logic for determining these setpoints is disclosed in Figure 2. The inner loop in Figure 2 runs for the duration of an entire market interval, after which it is time to compute a reward r, which specifies how good the action a was with respect to the multi-objective optimization targets stated under the purpose of this invention. The notation [] is appended to any variable defined in this invention to indicate an array storing the values for each timestep in the recent market interval. Thus, such an array has N elements. r is defined as follows:

$$r = rev() + c_1 pen() \tag{9}$$

where c_1 is the weight used to adjust the importance of penalty in the multi-objective optimization.

$$rev() = e_{sold} * price_{accept} \tag{10}$$

pen() is the possible penalties from failing to deliver the sold electricity and is a function of e_{sold} and $e_{delivered}$. *pen()* depends on the market rules of the specific electrical power market. The goal is to maximize r by maximizing *rev()* and minimizing *pen()*. Thus, the weight c_1 should be set so that the last term of equation (9) is negative.

The method for training the reinforcement learning agent in Figure 2 will require a number of iterations, before the agent will be able to bid successfully. Thus, it is advantageous to perform the training with a simulation rather than with the physical equipment, also using a market simulator developed with historical market price data. After a training phase in the simulated environment, the agent can be deployed to the physical environment. A properly trained reinforcement learning agent will be able to generalize, so it can perform well even if the simulation is not a high-fidelity model of the physical environment. The agent can continue to learn as it continues to execute the method in Figure 2, as it bids in the physical environment.

The action selected by the reinforcement learning agent will impact the SoC at the beginning of the next market interval. Thus, it will also impact the reward received for whatever action is taken during the next market interval. To account for such impacts, reinforcement learning employs a discounting technique to weight estimated future rewards, so that the weights decrease for rewards further in the future, as they are more uncertain.

IV. Implementation

The environment for the reinforcement learning agent is the implementation of the flowchart in Figure 2 for Finnish spot markets. The Finnish spot market is a day-a-head auction type market [19]. The implementation follows the Open AI Gym interface with callable functions *step()* and *reset()* [20]. The reset function returns the observation and initializes the battery SoC to random integer between 5 and 95. The *step()* returns reward, done and observation (i.e., the next state *s*). The implementation is in python, except the actions after the False branches of decisions 102, 105 and 106.

The battery management system and battery storage are implemented with MATLAB Simulink that is called from python. The battery implementation follows the basic principles presented in [9]. The implementation is presented in Figure 3. The generic battery model [21] is controlled with current source and the current is calculated from equation:

$$i(t)=P(t)/U(t) \tag{11}$$

where $P(t)$ is the constant power and $U(t)$ is the SoC dependent battery voltage. The battery voltage is limited with a saturation block between fully charged and cut-off voltage. As a result, the simulation returns the SoC and in case of a fully discharged battery, the amount of undelivered energy.

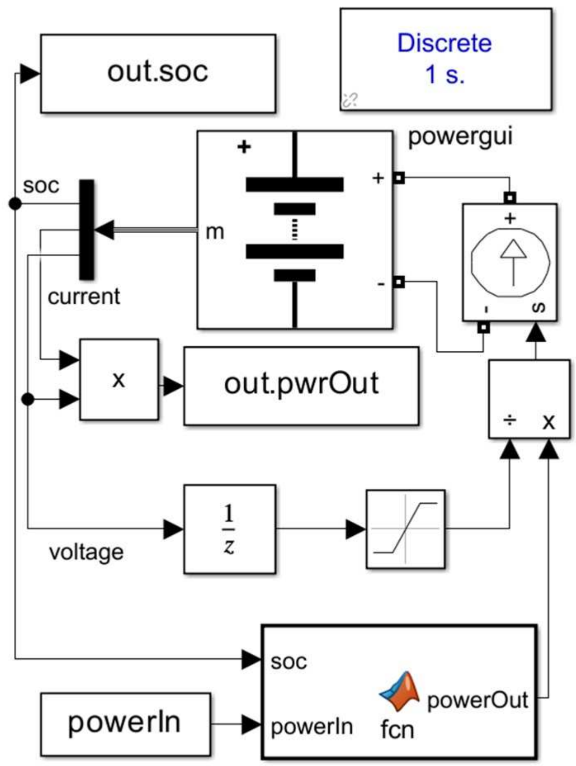

Fig. 3. Battery storage and management system

The relevant parameters of the battery are presented in Table I. The discharging characteristics are determined from the nominal parameters of the battery.

TABLE I. BATTERY PARAMETERS

Parameter	Value
Nominal Voltage	1500 V
Rated Capacity	1200 Ah
SoC limits	[5,95]
Battery response time	3 s
Sample time	1 s

The reinforcement learning agent is using Proximal Policy Optimization (PPO) [22] as a learning algorithm. In comparison to other RL-based algorithms like Deep Q Network (DQN), PPO is more stable, performs better, and has a higher rate of convergence. Also, PPO is simple to adopt, and easy to get working. The PPO algorithm is implemented with Stable-baselines3 (PyTorch) [23]. The initial parameters of the PPO solution are documented in [22]. The relevant parameters for the agent are in Table II:

TABLE II. PPO AGENT PARAMETERS

Parameter	Value
Learning rate, γ	0.01
Horizon	216
Minibatch size	24
Number of epochs	9
Rollout buffer size	216
GAE parameter	0.95
Number of hidden layers	2
Numbers of nodes in input layer	14
Number of nodes in output layer	2
Number of nodes in hidden layers	64
Activation functions	tanh

The agent uses the history data of predictions, price information, hours of sunrise and sunset and production of PV panels from Marjamäki industrial estate by the Energy Self-Sufficient Lempäälä. The time range of 08-May-2020 to 23-Dec-2020 was used for training and validation. The data for validation was 10% of the total dataset and was randomly selected.

V. Results

The smoothed (by exponential moving average) reward, with smoothing weight of 0.6 is presented in Figure 4. From the figure it is observable that the agent's learning curve is increasing heavily until around 200000 steps, where one hour is one step, after which the reward shows only minor improvement. The major grids are presented every 20 epochs where one epoch is the time where the agent has seen the full training data once. The input states vary from epoch to another as the battery usage from past ten hours is used as an input state. Also, the battery state of charge is randomized before every day.

The cumulative reward of 23 validation days is presented in Figure 5. From the Figure, it is observable that in general the validation rewards follow the trend of the training data.

The best candidate for the agent, based on validation data, is the agent after 93 epochs. In Figure 6 presents the bidding behavior for one example day: 01-Oct-2020. In Figure 6.a is the predicted PV production with a binary filter of daylight.

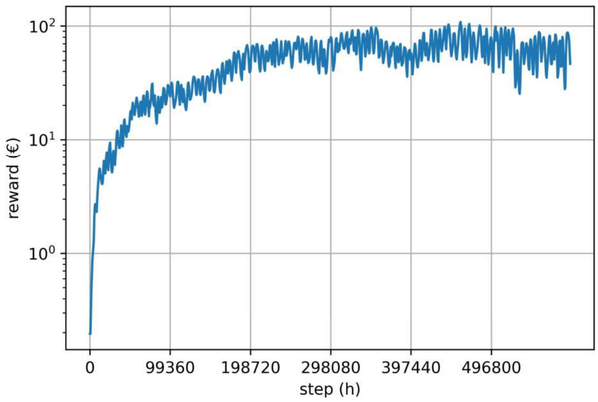

Fig. 4. Exponential moving average of the reward

In Figure 6.b is the bid from the agent. In Figure 6.c is the predicted market price in euros. Figure 6.d includes the predictions of relative price and relative PV production (Equations 7 and 8).

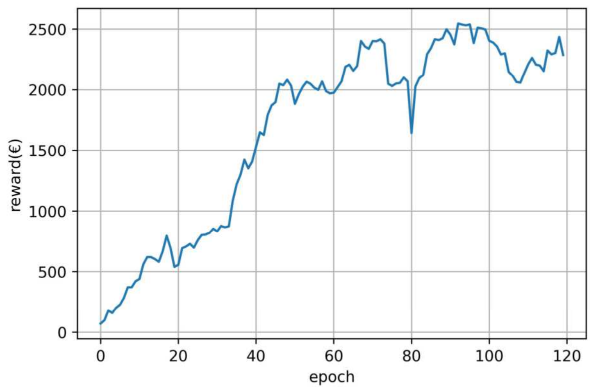

Fig. 5. Validation cumulative reward of 23 days due to the agent's actions

From Figure 6 it is observable that the bid is dependent on the combination of input states. When the predictions of PV productions are zero, the bid is in general zero. The exception for this is when predicted price and relative price prediction are high (Hours 4-5). When predictions of PV production (absolute and relative) are low, the bid goes up a bit. This means that the agent uses the battery to gain the advantage of market price. When the Prices drop, the bid also drops a bit. When predictions of PV production start to increase, the bids start to grow. The bids stay around 200 kW as long as there are high enough predictions for PV production and prices. When the predictions go under some threshold (Hours 17-18), the bids go to zero.

VI. Conclusions

In conclusion, integrating batteries into renewable energy systems is a challenging process that calls for a thorough analysis of the market environment, system performance, and battery SoC. The best bids must consider not just the current market interval but also how the bids might affect the system's performance throughout several consecutive market intervals, especially due to the possibility of the battery becoming full or empty. The renewable energy system's profitability can be increased through this optimization process, enabling it to take full advantage of any potential market opportunities.

Acknowledgment

This research was funded by Business Finland grant number 1516/31/2022.

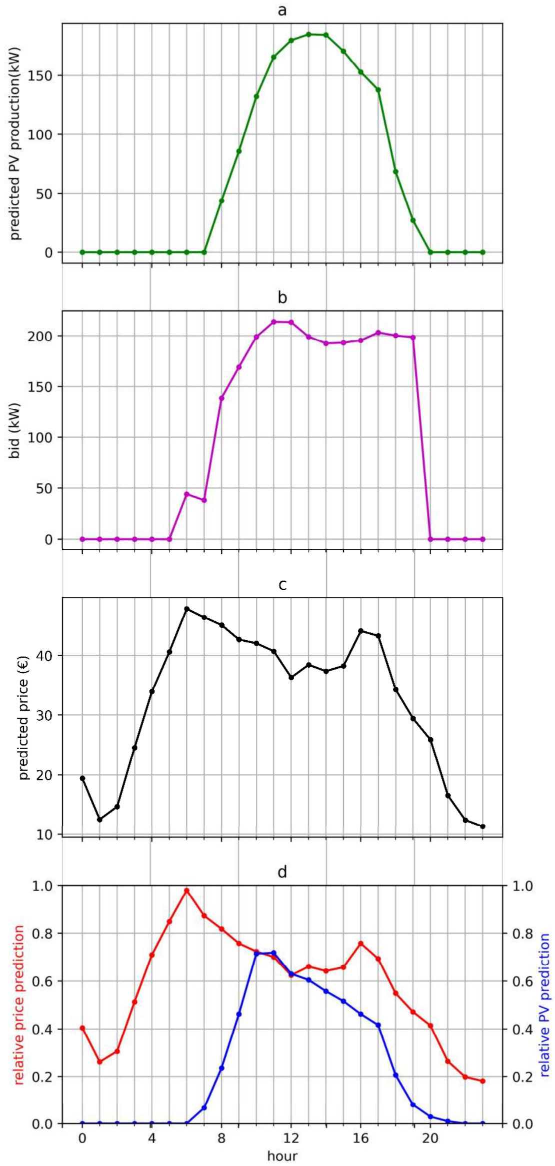

Fig. 6. Predicted price and the bids

References

[1] T. Chen, Q. Hu, Y. Wang, X. Quan, F. Zhang, and X. Yan, "Residential Energy Arbitrage Considering Demand Response and Accurate Battery Model with Degradation Cost," in *2022 International Conference on Power Energy Systems and Applications (ICoPESA)*, Feb. 2022, pp. 511–515, doi: 10.1109/ICoPESA54515.2022.9754456.

[2] R. K. Lam, D. H. Tran, and H.-G. Yeh, "Economics of residential energy arbitrage in california using a PV system with directly connected energy storage," in *2015 IEEE Green Energy and Systems Conference (IGESC)*, Nov. 2015, pp. 67–79, doi: 10.1109/IGESC.2015.7359453.

[3] K. Zeng, J. Liu, H. Wang, Y. Zhou, and S. Zhou, “Supportive Analysis of Battery Energy Storage System on Electricity Market Scheduling Considering Load and Wind Power Fluctuation,” in *2021 IEEE 4th International Electrical and Energy Conference (CIEEC)*, May 2021, pp. 1–4, doi: 10.1109/CIEEC50170.2021.9510779.

[4] B. Zakeri and S. Syri, “Economy of electricity storage in the Nordic electricity market: The case for Finland,” in *11th International Conference on the European Energy Market (EEM14)*, May 2014, pp. 1–6, doi: 10.1109/EEM.2014.6861293.

[5] S. Sano, Y. Ito, Y. Hirose, H. Takeuchi, and S. Aone, “Development of long cycle life valve-regulated lead-acid battery for large-scale battery energy storage system to utilize renewable energy,” in *2015 IEEE International Telecommunications Energy Conference (INTELEC)*, Oct. 2015, pp. 1–6, doi: 10.1109/INTLEC.2015.7572367.

[6] R. Subramanya, S. Sierla, and V. Vyatkin, “From DevOps to MLOps: Overview and Application to Electricity Market Forecasting,” *Appl. Sci.*, vol. 12, no. 19, p. 9851, Sep. 2022, doi: 10.3390/app12199851.

[7] M. Huang, X. Wang, and S. Zhang, “Analysis of an electricity market equilibrium model with penalties for wind power’s bidding deviation,” in *2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT)*, Nov. 2015, pp. 35–40, doi: 10.1109/DRPT.2015.7432216.

[8] H. Beltran, J. Barahona, R. Vidal, J. C. Alfonso, C. Arino, and E. Perez, “Ageing of different types of batteries when enabling a PV power plant to enter electricity markets,” in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, Oct. 2016, pp. 1986–1991, doi: 10.1109/IECON.2016.7794082.

[9] H. Aaltonen, S. Sierla, V. Kyrki, M. Pourakbari-Kasmaei, and V. Vyatkin, “Bidding a Battery on Electricity Markets and Minimizing Battery Aging Costs: A Reinforcement Learning Approach,” *Energies*, vol. 15, no. 14, p. 4960, Jul. 2022, doi: 10.3390/en15144960.

[10] W. Chen, X. Qiu, T. Cai, H.-N. Dai, Z. Zheng, and Y. Zhang, “Deep Reinforcement Learning for Internet of Things: A Comprehensive Survey,” *IEEE Commun. Surv. Tutorials*, vol. 23, no. 3, pp. 1659–1692, 2021, doi: 10.1109/COMST.2021.3073036.

[11] Y. Dong, T. Zhao, and Z. Ding, “An Optimal Day-ahead Bidding Strategy and Operation for Battery Energy Storage System by Reinforcement Learning,” *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 13190–13195, 2020, doi: 10.1016/j.ifacol.2020.12.144.

[12] D. Haputhanthri *et al.*, “Solar Irradiance Nowcasting for Virtual Power Plants Using Multimodal Long Short-Term Memory Networks,” *Front. Energy Res.*, vol. 9, Aug. 2021, doi: 10.3389/fenrg.2021.722212.

[13] R. Subramanya, M. Yli-Ojanperä, S. Sierla, T. Hölttä, J. Valtakari, and V. Vyatkin, “A Virtual Power Plant Solution for Aggregating Photovoltaic Systems and Other Distributed Energy Resources for Northern European Primary Frequency Reserves,” *Energies*, vol. 14, no. 5, 2021, doi: 10.3390/en14051242.

[14] T. Kempitiya, S. Sierla, D. De Silva, M. Yli-Ojanperä, D. Alahakoon, and V. Vyatkin, “An Artificial Intelligence framework for bidding optimization with uncertainty in multiple frequency reserve markets,” *Appl. Energy*, vol. 280, 2020, doi: 10.1016/j.apenergy.2020.115918.

[15] S. Kahawala *et al.*, “Robust Multi-Step Predictor for Electricity Markets with Real-Time Pricing,” *Energies*, vol. 14, no. 14, p. 4378, Jul. 2021, doi: 10.3390/en14144378.

[16] R. A. Chinnathambi and P. Ranganathan, “Investigation of forecasting methods for the hourly spot price of the day-ahead electric power markets,” in *2016 IEEE International Conference on Big Data (Big Data)*, Dec. 2016, pp. 3079–3086, doi: 10.1109/BigData.2016.7840962.

[17] Y. Bai, J. Wang, and W. He, “Energy arbitrage optimization of lithium-ion battery considering short-term revenue and long-term battery life loss,” *Energy Reports*, vol. 8, pp. 364–371, Dec. 2022, doi: 10.1016/j.egyr.2022.10.209.

[18] P. L. C. García-Miguel, A. P. Asensio, J. L. Merino, and M. G. Plaza, “Analysis of cost of use modelling impact on a battery energy storage system providing arbitrage service,” *J. Energy Storage*, vol. 50, p. 104203, Jun. 2022, doi: 10.1016/j.est.2022.104203.

[19] E. A. Unger, G. F. Ulfarsson, S. M. Gardarsson, and T. Matthiasson, “A long-term analysis studying the effect of changes in the Nordic electricity supply on Danish and Finnish electricity prices,” *Econ. Anal. Policy*, vol. 56, pp. 37–50, Dec. 2017, doi: 10.1016/j.eap.2017.06.001.

[20] G. Brockman *et al.*, “OpenAI Gym,” Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.01540.

[21] Mathworks, “Generic Battery Model,” 2023. https://www.mathworks.com/help/sps/powersys/ref/battery.html (accessed Feb. 26, 2023).

[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.06347.

[23] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, “Stable-baselines3: Reliable reinforcement learning implementations,” *J. Mach. Learn. Res.*, vol. 22, 2021.