Risk-sensitive Reinforcement Learning-based Strategies for Dutch Implicit Balancing

Seyed Soroush Karimi Madahi
IDLab, Ghent University – imec
Ghent, Belgium
seyedsoroush.karimimadahi@ugent.be

Fabio Pavirani
IDLab, Ghent University – imec
Ghent, Belgium
fabio.pavirani@ugent.be

Bert Claessens
Beebop
Belgium
bert@beebop.ai

Chris Develder
IDLab, Ghent University – imec
Ghent, Belgium
chris.develder@ugent.be

***Abstract*—Adopting renewable energy sources (RES) can pave the way toward reaching net-zero carbon emissions. However, the intermittent nature of RES can pose significant challenges to balance responsible parties (BRPs) and transmission system operators (TSOs) in maintaining the balance of the electricity grid. BRPs can assist TSOs in balancing the grid by occasionally deviating from their nomination to help reduce the system imbalance, which is called implicit balancing. In this paper, we propose data-driven implicit balancing strategies for BRPs in the Dutch imbalance settlement mechanism. Dutch implicit balancing is challenging due to the Dutch imbalance pricing calculation, which is a combination of single and dual pricing methods. To cope with this challenge, a risk management perspective is incorporated into the proposed method through distributional reinforcement learning. Distributional reinforcement learning agents are trained to manage a BRP's battery in the presence of wind farm generation stochasticity to reduce its imbalance cost. Dutch imbalance data of 2024 are used to assess the performance of the learned implicit strategies. The proposed method is benchmarked against deterministic model predictive control and a rule-based controller. The results show that both risk-neutral and risk-averse agents improve daily profit by 29.3% and 20.7%, respectively, compared to the rule-based controller. Moreover, the risk-averse agent decreases the average portfolio deviation during dual pricing situations by 19.2% compared to the risk-neutral agent, resulting in a lower imbalance cost for the BRP in these situations.**

***Index Terms*—Battery, Imbalance settlement mechanism, Implicit balancing, Reinforcement learning, Risk-sensitive arbitrage**

I. Introduction

To achieve net zero carbon emissions, the share of renewable energy sources (RES) is steadily increasing in the current energy mix. Yet, the intermittent nature of RES generation causes greater uncertainty in the system imbalance. To help maintain that grid balance, transmission system operators (TSOs) outsource part of the required corrective balancing actions to balance responsible parties (BRPs). At the end of each imbalance settlement period (ISP), unbalanced BRPs are exposed to an imbalance price, calculated based on the total system imbalance to penalize deviations against the system balance. In certain European countries, e.g., Belgium and the Netherlands, BRPs can also be remunerated if their unbalanced portfolio contributes to restoring grid balance, i.e., if their deviation is opposite to the direction of the system imbalance. The European electricity balancing guideline states that the pricing method in such implicit balancing should financially incentivize these BRPs, while penalizing the rest [1]. However, performing implicit balancing is challenging, as it involves risks mainly stemming from the volatility of imbalance prices and prediction errors in RES production.

979-8-3315-1278-1/25/$31.00 © 2025 IEEE

Various methods have been used in previous research works for implicit balancing. In [2], a stochastic model predictive control (MPC) method was proposed to optimize the profit of batteries in the Belgian imbalance settlement mechanism by considering battery degradation costs and risk aversion. The authors in [3] proposed a novel data-driven inverse optimization method to model demand response resources using their aggregate forecasts. The proposed method was integrated into the Belgian imbalance settlement mechanism to optimize the imbalance cost of BRPs. In [4], a mixed-integer linear programming (MILP) optimization was used to obtain the optimal trading strategies for both the Dutch day-ahead market and Dutch imbalance settlement mechanism. All the aforementioned works [2]–[4] used model-based optimization methods for implicit balancing. These methods require a sufficiently accurate model of the system, as well as an imbalance price forecaster. However, forecasting imbalance prices is challenging because of their high uncertainty [5], [6]. Furthermore, these methods suffer from a high computational burden during inference due to the repeated solving of an optimization problem. This limits their application to problems with relatively short decision-making intervals, such as a minute-based implicit balancing problem.

Few articles have explored data-driven methods for implicit balancing. MPC-based and reinforcement learning (RL)-based controllers were developed in [7] for seasonal thermal energy storage systems to interact with the day-ahead and imbalance markets in Belgium and the Netherlands. However, they did not study the problem from a risk-sensitive perspective, and their decision-making time resolution is 15 minutes. In [8], an RL-based battery control framework was proposed for risk-sensitive energy arbitrage in the Belgian imbalance settlement mechanism, considering a cycle constraint. Although [2], [8] introduced risk-sensitive strategies, they focused on the Belgian imbalance settlement mechanism, which uses the single-pricing method. Additionally, they did not consider the effect of RES generation uncertainty on their risk-sensitive strategies.

Among the aforementioned studies, only [4], [7] has focused on implicit balancing strategies for the Dutch imbalance settlement mechanism. Since the Dutch imbalance pricing method is a mixture of single and dual pricing, implicit balancing in the Dutch imbalance settlement mechanism is more challenging than in other imbalance settlement mechanisms. Dutch implicit balancing strategies must handle uncertainty in dual pricing occurrence, along with price and RES production uncertainties. Furthermore, risk management plays an important role in the Dutch imbalance settlement due to the occurrence of situations with dual imbalance pricing. In such situations, as all unbalanced portfolios are penalized, BRPs must keep their portfolios balanced to avoid incurring imbalance costs, which becomes particularly challenging if their portfolios include any type of RES. Considering the stochasticity of RES generation and uncertainty in the occurrence of dual pricing, BRPs need to adopt a risk-averse implicit balancing strategy to mitigate the risk of incurring significant imbalance costs during dual pricing situations, while ensuring profitability in single pricing situations. None of the aforementioned research investigated the effect of RES production uncertainty on implicit balancing strategies.

To address the aforementioned gaps in the literature — (i) rare research on data-driven alternatives for Dutch implicit balancing, (ii) a lack of investigation into a risk-sensitive perspective in the Dutch imbalance settlement mechanism, and (iii) limited studies on the effect of RES generation stochasticity on implicit balancing strategies — we propose a data-driven risk-sensitive implicit balancing strategy for managing a BRP's portfolio, which includes a battery and a wind farm exposed to the Dutch imbalance settlement mechanism. We train an RL-based agent to control the battery minute by minute in order to minimize the BRP's imbalance cost. We focus on distributional RL due to its outstanding performance compared to standard RL [9]. Moreover, distributional RL has the ability to learn the complete probability distribution of returns, rather than a single value for the expected return, which makes it suitable for risk management. The performance of the proposed method is evaluated on the Dutch imbalance data of 2024, which consist of both single and dual pricing. Our main contributions in this paper can be summarized as follows:

- Propose **risk-sensitive** implicit balancing strategies to manage the portfolios of BRPs, exemplified in a case study considering a battery and wind farm;
- Control the battery every minute to perform implicit balancing in the **Dutch** imbalance settlement mechanism using **distributional RL** agents with a continuous action space;
- Benchmark the proposed data-driven method against the deterministic MPC and a rudimentary rule-based controller (RBC).

II. Problem Formulation

A. Dutch Imbalance Settlement Mechanism

TSOs are responsible to maintain the grid frequency at 50 Hz. To fulfill this responsibility, TSOs ask BRPs to submit their nomination the day before delivery. If BRPs deviate from their nomination during the delivery day, they create an imbalance. BRPs are financially responsible for the imbalance they cause during the ISP (15 min in most European countries). TSOs continuously monitor the total system imbalance and grid frequency. Any change in the frequency will be compensated for by activating the appropriate reserve capacity offered by balancing service providers. At the end of the ISP, the imbalance settlement prices are determined and unbalanced BRPs are invoiced. To encourage BRPs to restore grid balance during the ISP, the settlement mechanism is designed to be costly for deviations that increase the system imbalance and profitable for deviations that reduce it.

In the Netherlands, the imbalance price for each ISP is calculated based on the frequency restoration reserve (FRR) volume activated by the Dutch TSO (TenneT) and the activation situation of balancing energy (referred to as the regulation state). Regulation state 0 applies to a situation in which TenneT does not activate any upward or downward FRR during the ISP. In this situation, the imbalance price is set to the mid-price, defined as the average between the lowest upward bid price and the highest downward bid price in the merit order. Regulation state 1 means that TenneT either activates upward FRR or both upward and downward FRR directions, with continuously increasing balance deltas (activated upward FRR volume minus activated downward FRR volume) within the ISP. In this case, the price equals the highest price of the activated upward aFRR/mFRR. Regulation state -1 indicates that TenneT either activates downward FRR or both upward and downward FRR directions, with continuously decreasing balance deltas throughout the ISP. The price is equal to the lowest price of the activated downward aFRR/mFRR. Finally, if TenneT activates both FRR directions without monotonic balance deltas within the ISP, regulation state 2 occurs, leading to the dual pricing settlement [10]. In regulation state 2, a BRP shortage/surplus is exposed to the marginal upward/downward price. In other words, all unbalanced BRPs will be penalized, regardless of their direction. This pushes BRPs to keep their portfolios balanced to avoid incurring imbalance costs.

Due to a mixture of single and dual pricing, implicit balancing in the Dutch imbalance settlement mechanism is more complex compared to other imbalance settlement mechanisms, such as the Belgian one (which uses a fully single pricing method [11]) and the Portuguese one (which uses a fully dual pricing method [12]). In the Dutch settlement mechanism, the proposed implicit balancing strategy needs to deal with regulation state uncertainty, as well as imbalance price volatility and RES stochasticity.

B. MDP Formulation

The implicit balancing strategies we propose aim to optimize the imbalance cost of the BRP by controlling battery actions. To this end, we model the problem as a Markov decision process (MDP) and solve it as a stochastic sequential decision-making problem. The MDP problem is formulated by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, Where $\mathcal{S}$ and $\mathcal{A}$ are the state and

action spaces, respectively, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines the reward function, $\mathcal{P} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the unknown state transition probability distribution, and $\gamma \in (0, 1]$ is the discount factor [13].

The state at each 1-minute time step t is defined as

$$s_t = (T_{\text{qh}}, \text{qh}, \text{mo}, \text{SOC}_t, \hat{\pi}_t^+, \hat{\pi}_t^-, \text{flag}_{\text{reg2}}, \omega_t, \text{dev}_{p,t}) \quad , \quad (1)$$

where $T_{\text{qh}} \in [0, 14]$, $\text{qh} \in [0, 95]$, and $\text{mo} \in \{1, 2, \ldots, 12\}$ denote the minute of the quarter hour, the quarter hour of the day, and the month of the year, respectively, SOC_t is the state of charge (SoC) of battery at time t, and $\text{dev}_{p,t}$ represents the total portfolio energy deviation from its nomination at time t. Additionally, ω_t denotes the wind power deviation of the BRP from its nomination, i.e., the wind generation forecast error of the BRP. $\text{flag}_{\text{reg2}}$ is a boolean flag that indicates whether the grid is in regulation state 2 (dual pricing mode). $\hat{\pi}_t^+$ and $\hat{\pi}_t^-$ are the *indicative* positive and negative imbalance prices for the current quarter hour qh, respectively. Indicative prices are minute-based prices published by TenneT in real-time, based on the latest FRR activations and grid situation, providing more information to BRPs. Since the actual imbalance prices are calculated ex-post, our agent can only make decisions based on these indicative imbalance prices.

The agent's action is continuous, given by

$$a_t \in \mathcal{A}, \quad \mathcal{A} = [-P_{\text{max}}, P_{\text{max}}] \quad , \quad (2)$$

where P_{max} represents the maximum (dis-)charging power of the battery, and a positive value of a_t indicates charging the battery. In this paper, the time granularity for decision-making is 1 minute. The intuition behind it is that, as the end of the quarter hour approaches, the agent has access to more accurate information about the grid situation. Therefore, the agent can continuously adjust its position to align with the correct direction of the system, rather than keeping it fixed for the whole ISP.

Since the RL agent aims to maximize arbitrage revenue, the reward function is defined as the negative of the imbalance cost, as below:

$$r_t = \begin{cases} -(\omega_t - a_t)\pi_{\text{qh}}^{\text{imb}+} & : \sum_{\text{t} \in \text{qh}} \text{dev}_{p,t} > 0 \\ -(\omega_t - a_t)\pi_{\text{qh}}^{\text{imb}-} & : \sum_{\text{t} \in \text{qh}} \text{dev}_{p,t} \leq 0 \end{cases} \quad , \quad (3)$$

where $\pi_{\text{qh}}^{\text{imb+}}$ and $\pi_{\text{qh}}^{\text{imb}-}$ denote the real positive and negative imbalance prices of the quarter hour in which t lies, respectively. Note that in the single pricing situation, the positive and negative prices are identical.

System dynamics are modeled using a state transition probability function $\mathcal{P}$. The battery dynamics form part of the system dynamics in our problem, which we explicitly formulate using a battery linear model with a constant round-trip efficiency for (dis)charging, as in [14]. However, $\mathcal{P}$ is primarily unknown, due to dependency on uncertainties in imbalance prices, regulation state, weather, and wind forecast error. By interacting with the environment, the RL agent can implicitly learn $\mathcal{P}$ and these uncertainties.

C. Wind Power Forecast Error Scenario Generation

As mentioned earlier, the imbalance cost and portfolio balance heavily depend on the wind power forecast error. Thus, we need to evaluate the robustness of our implicit balancing strategies under realistic forecast error scenarios. To generate the scenarios, we employ an experimental distribution-based method introduced in [15]. In this method, scenarios are sampled from historical distributions for the forecast error. Inter-temporal correlations between different quarter hours are captured through the covariance matrix calculated using historical data. We use an extended version of this method, where scenarios are generated using conditional distributions based on the forecast error of previous quarter hours. A more detailed explanation of the method can be found in [16].

III. Distributional SAC

The MDP problem introduced in Section II-B is solved using RL. RL aims to learn a policy that maximizes the expected long-term reward. We focus on policy gradient RL methods because they support continuous actions. Among existing policy gradient methods, we choose soft actor-critic (SAC) for its superior sample efficiency and stability [17].

In distributional RL, the probability distribution over returns is approximated instead of estimating an expected return. Incorporating a distributional perspective into RL provides various advantages, such as supporting the learning of risk-sensitive policies, alleviating Q-value overestimation, and enhancing the stability of training. We use quantile regression to estimate the return distribution.

To learn a risk-sensitive policy, we replace the expectation operator in the Bellman equation with a risk measure function. In this paper, we apply Value-at-Risk (VaR) with a confidence level of 0.1. In this way, the agent is trained to maximize the lower 10% tail of the return distribution. In other words, the agent learns a policy that mitigates the distribution tail. The mathematical formulation of distributional SAC and incorporating the risk measure is outlined in detail in appendix A.

IV. Simulation Results

A. Simulation Setup

We study the effectiveness of our proposed implicit balancing methods using the Dutch imbalance prices of 2024. Since the final imbalance price is calculated based on the most extreme FRR activated during the quarter hour, the decision-making time resolution must be adequately short to take advantage of the grid's most recent state, which is set to 1 minute in this paper. We consider a BRP with a 10 MW wind farm and a 1 MW/2 MWh battery with 90% round-trip efficiency and a minimum SoC of 10%.

The dataset is split such that the first 20 days of each month form the training set, the 21st to 25th form the validation set, and the rest form the test set. The RL agents are trained over $50\,000$ episodes, with each episode representing a single day. The experience replay buffer size, the mini-batch size, the soft update factor τ, and the discount factor γ, are equal to 1×10^6, $16\,384$, 0.1, and 0.9995, respectively. The actor and

critic networks are modeled as fully connected neural networks with learning rates of 5×10^{-5} and 5×10^{-4}, respectively.

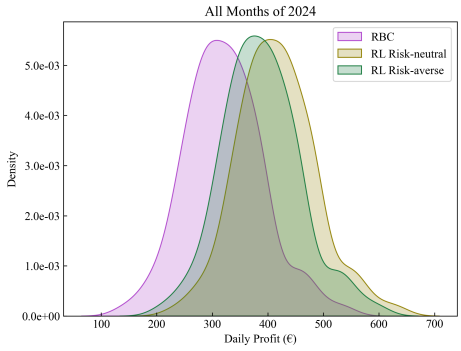
To benchmark the proposed implicit balancing strategies, a rule-based controller and MPC are introduced. The RBC control logic in this paper is as follows: If the grid is in a single-pricing situation, the battery action is determined based on fixed thresholds for (dis)charging, which are set to the first and third quartiles of the 2024 prices—31.9 €/MWh and 94.9 €/MWh, respectively. In a dual-pricing situation, the battery tries to maintain the balance of the portfolio. In the deterministic MPC, we formulate the arbitrage problem as a MILP optimization problem with perfect foresight for different look-ahead horizons. We carry out two different experiments to demonstrate the performance of various methods. The first experiment focuses on the robustness of agents against various generated scenarios of the wind forecast error. In the second experiment, we study their performance on the historical data from the test days.

B. Results

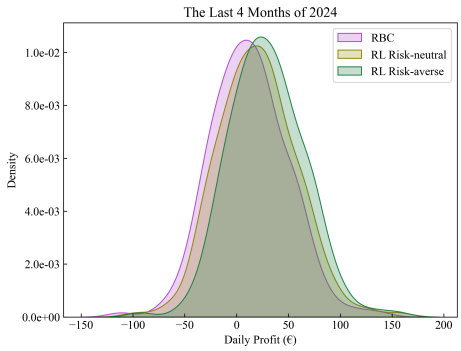
We evaluate the performance of the implicit balancing strategies on the test days using 200 generated scenarios for wind production forecast errors, as explained in Section II-C. Figure 1a depicts the daily profit distribution of the methods. Both risk-neutral and risk-averse agents significantly outperformed RBC in terms of average daily revenue by 29.3% and 20.7%, respectively. The average portfolio deviation during dual pricing instances is directly linked to the risk-aversion level of the methods, shown in Fig. 2: The lower the absolute value, the better the agent is at hedging against risk. Note that a negative portfolio deviation indicates that the BRP is short, meaning its energy consumption exceeds its nomination. The risk-averse RL agent decreased the average portfolio deviation by 19.2% compared to that of the risk-neutral one. RBC performed better than RL at balancing the portfolio because of its cautious bounds for (dis)charging.

To make the experiment riskier, we consider only test days from the last 4 months of 2024, when the occurrence of regulation state 2 raised from 11.9% in the first 8 months to 24.1%. As Fig. 1b shows, the risk-averse agent achieved higher daily revenues compared to the other two agents. The risk-averse agent boosts the average daily profit from 37.3€ for the RBC and 60.4€ for the risk-neutral one to 88.1€. The VaR value for the RBC, risk-neutral, and risk-averse agents is -92€, -75.8€, and -41€, respectively. The risk-averse revenue distribution has the highest VaR value, which aligns with its actor network loss function (Eq. (12)). Figure 2b illustrates that the risk-neutral agent causes a larger portfolio deviation on average compared to the risk-averse agent, leading to greater penalties for the BRP.

To better interpret the policies learned by the RL agents, in Fig. 3, we illustrate their policy heat map with respect to imbalance price and SoC, the two most important features for the agents. The figures highlight that both agents are sensitive to portfolio deviation, even in single-pricing situations. The reason is that, although the grid operates in single-pricing



(a)

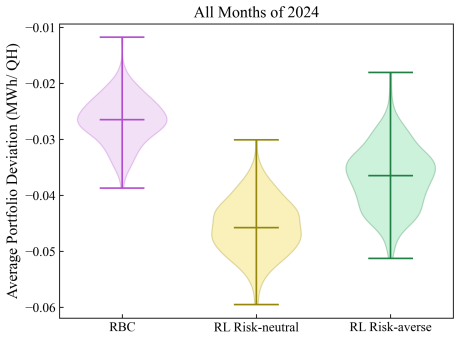


(b)

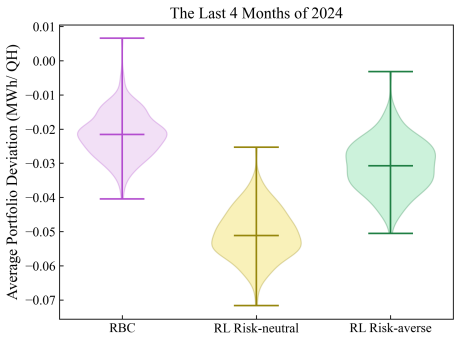
Figure 1: The probability distribution of daily profit for 200 generated scenarios.

mode in real-time, there is still a chance that it shifts to dual-pricing mode. Therefore, the agents must adjust their policies based on the current portfolio deviation. Moving from BRP shortage to BRP surplus, the charge area expands and the discharge area is restricted in all the figures. A comparison of the learned policies for single pricing shows that the risk-averse agent has a larger charge area for the same portfolio deviation as the risk-neutral one. This is due to the risk-averse nature of the agent, which ensures there is always enough energy in the battery for the future. In the dual-pricing situation, current portfolio deviation becomes more dominant and both agents change their policy to focus on offsetting the portfolio energy deviation.

Finally, Table I indicates the performance of the different methods on the test days using historical real wind forecast error data. The perfect foresight MPC results demonstrate the upper-bound performance for RBC and RL. Using MPC results in the least portfolio deviation because it can effectively plan to compensate for deviations in dual pricing situations, given its access to accurate future information (including wind forecast errors). Both risk-neutral and risk-averse RL agents achieved higher performance than RBC, with daily revenues increasing by 16.2% and 11.9%, respectively. The risk-averse RL agent achieved a 52.6% reduction in the average portfolio deviation compared to the risk-neutral agent, even though it



(a)



(b)

Figure 2: The portfolio deviation distribution for 200 generated scenarios.

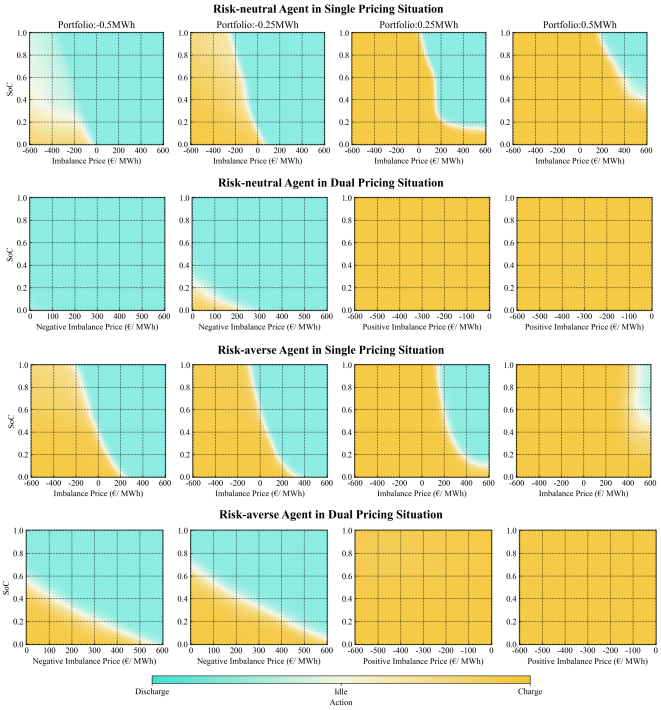


Figure 3: The learned policies for risk-neutral and risk-averse RL agents

Table I: Evaluation of the methods on the historical data from the test days

Method	Avg. Daily Revenue (€)	Avg. Portfolio Dev.* (MWh/ QH)
MPC horizon 15 minutes	825.79	0
MPC horizon 1 hour	1096.67	0
MPC horizon 4 hours	1235.46	0
MPC horizon 10 hours	1295	0
RBC	529.49	-1.33×10^{-3}
RL (risk-neutral)	615.22	-25.62×10^{-3}
RL (risk-averse)	592.44	-12.14×10^{-3}

* The average portfolio deviation is calculated over quarter hours with regulation state 2, as in the other situations, deviation in the right direction is beneficial.

slightly decreased the daily profit.

V. Conclusion

In this paper, we proposed an RL-based risk-sensitive method for Dutch implicit balancing to manage the portfolio of BRPs, consisting of a battery and wind farm. We validated the performance of the proposed method using the Dutch imbalance data from 2024. The results showed that on days with fewer dual pricing occurrences, the risk-neutral agent outperformed the risk-averse agent and RBC. However, the risk-averse agent reduced the average portfolio deviation by 19.2% by learning a policy with a larger charge area and maintaining the SoC at a sufficient level for the next quarter hours. On the other hand, for riskier days with more frequent instances of regulation state 2, our experiment revealed that the risk-averse agent achieved higher average daily revenue compared to other agents while hedging against the risk of incurring large daily imbalance costs. The risk-averse agent improved the average daily profit of the risk-neutral agent and RBC by 45.86% and 136.2%, respectively. The robustness experiments highlight an importance of adopting the risk-averse strategy for Dutch implicit balancing, especially nowadays due to continuous increase in regulation state 2 instances.

In future research, we will focus on incorporating more grid-related inputs into the agents to inform them about the most recent grid condition. Adding features such as minute-based activated FRR volumes can help the agents better capture the possibility of regulation state 2 occurrence in the following minutes. Another direction for future work is to incorporate the battery degradation cost into the reward function. Frequently alternating between charging and discharging can significantly reduce the battery's lifespan, resulting in a decrease in net profit.

Acknowledgment

This research was partly funded by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme and the Horizon Europe project BlueBird (https://bluebird-project.eu/ - grant agreement no. 101192452).

REFERENCES

[1] European Network of Transmission System Operators for Electricity, "Explanatory document to all tsos' proposal to further specify and harmonise imbalance settlement in accordance with article 52(2) of commission regulation (eu) 2017/2195 of 23 november 2017, establishing a guideline on electricity balancing," 2018.

[2] R. Smets, K. Bruninx, J. Bottieau, J.-F. Toubeau, and E. Delarue, "Strategic implicit balancing with energy storage systems via stochastic model predictive control," *IEEE Transactions on Energy Markets, Policy and Regulation*, vol. 1, no. 4, pp. 373–385, 2023.

[3] B. Vatandoust, B. B. Zad, F. Vallée, J.-F. Toubeau, and K. Bruninx, "Integrated forecasting and scheduling of implicit demand response in balancing markets using inverse optimization," in *2023 19th International Conference on the European Energy Market (EEM)*. IEEE, 2023, pp. 1–6.

[4] S. de Weerd, M. Gibescu, M. de Leeuw, J. van Haperen, and B. Siebenga, "Modelling the economic feasibility of distributed flexibility assets in the dutch electricity markets," in *2022 18th International Conference on the European Energy Market (EEM)*. IEEE, 2022, pp. 1–8.

[5] F. Pavirani, J. Van Gompel, S. S. K. Madahi, B. Claessens, and C. Develder, "Predicting and publishing accurate imbalance prices using monte carlo tree search," *arXiv preprint arXiv:2411.04011*, 2024.

[6] J. Van Gompel, B. Claessens, and C. Develder, "Probabilistic forecasting of power system imbalance using neural network-based ensembles," *arXiv preprint arXiv:2404.14836*, 2024.

[7] J. Lago, G. Suryanarayana, E. Sogancioglu, and B. De Schutter, "Optimal control strategies for seasonal thermal energy storage systems with market interaction," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 5, pp. 1891–1906, 2020.

[8] S. S. K. Madahi, B. Claessens, and C. Develder, "Distributional reinforcement learning-based energy arbitrage strategies in imbalance settlement mechanism," *Journal of Energy Storage*, vol. 104, p. 114377, 2024.

[9] J. Duan, Y. Guan, S. E. Li, Y. Ren, Q. Sun, and B. Cheng, "Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 11, pp. 6584–6598, 2021.

[10] TenneT, "Imbalance pricing system: how are the (directions of) payment determined?" pp. 1–17, 2022.

[11] J. Baetens, J. Laveyne, G. Van Eetvelde, and L. Vandevelde, "Imbalance pricing methodology in belgium: Implications for industrial consumers," in *2020 17th International Conference on the European Energy Market (EEM)*. IEEE, 2020, pp. 1–6.

[12] H. Algarvio, A. Couto, and A. Estanqueiro, "A double pricing and penalties "separated" imbalance settlement mechanism to incentive self balancing of market parties," in *2024 20th International Conference on the European Energy Market (EEM)*. IEEE, 2024, pp. 1–6.

[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 2018.

[14] S. s. Karimi madahi, G. Gokhale, M.-S. Verwee, B. Claessens, and C. Develder, "Control policy correction framework for reinforcement learning-based energy arbitrage strategies," in *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, 2024, pp. 123–133.

[15] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl, "From probabilistic forecasts to statistical scenarios of short-term wind power production," *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, vol. 12, no. 1, pp. 51–62, 2009.

[16] R. Smets, E. Delarue, K. Bruninx, and J.-F. Toubeau, "Participation of energy storage systems in short-term electricity markets: Exploring the interaction between optimization and machine learning," 2024.

[17] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[18] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *International conference on machine learning*. PMLR, 2017, pp. 449–458.

[19] W. Dabney, M. Rowland, M. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

APPENDIX

In SAC, an actor network π_ϕ is trained to yield the policy by maximizing the Q-values learned by a critic network π_ϕ, along with maximizing the entropy to encourage the agent to explore more. The loss function of the actor network (J_π) and the critic network (L_Q) are calculated as follows

$$J_\pi(\phi) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi}[\alpha \ln \pi_\phi(a|s) - Q_\theta(s,a)] \tag{4}$$

$$L_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}[(y_t - Q_\theta(s_t,a_t))^2] \tag{5}$$

$$y_t = r_t + \gamma \mathbb{E}_{a_{t+1}\sim\pi_\phi}[Q_{\theta'}(s_{t+1},a_{t+1}) - \alpha \ln \pi_\phi(a_{t+1}|s_{t+1})] \tag{6}$$

$$\theta' = \mu\theta + (1-\mu)\theta' \qquad \mu \ll 1 \tag{7}$$

In distributional RL, the distributional Bellman equation is used to estimate the distribution of returns [18]. We use quantile distribution to estimate the return distribution (Z_θ), parametrized as follows:

$$Z_\theta(s_t,a_t) = \sum_{i=1}^{N} q_{\tau_i}(s_t,a_t)\delta_{\tau_i} \quad , \tag{8}$$

where q_{τ_i} denotes the return value at the τ_ith quantile, and δ_{τ_i} is the Dirac delta function at τ_i. We consider N fixed quantiles as $\tau_i = \frac{i}{N}$. We set N to 20 in this paper. The main benefit of using quantile distribution rather than categorical distribution (introduced in [18]) is that the value distribution is not restricted to predefined bounds, which greatly improves prediction accuracy when the return range changes significantly across states [19].

By integrating the quantile regression perspective with SAC, the actor and critic loss functions (formulated in Eqs. (4) and (5)) are modified as below.

$$J_\pi(\phi) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi}\left[\alpha \ln \pi_\phi(a|s) - \mathbb{E}_{Z\sim Z_{\theta'}}[Z(s,a)]\right] \tag{9}$$

$$L_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}\left[\sum_{i=1}^{N} \mathbb{E}_j[(\tau_i - 1_{u_j<0})u_j]\right] \tag{10}$$

$$u_j = \mathcal{T}q_{\tau_j} - q_{\tau_i}(s_t,a_t) \tag{11}$$

To learn a risk-sensitive policy, we replace the expectation operator in Eq. (9) with a risk measure function (Ψ) as follows.

$$J_\pi(\phi) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi}[\alpha \ln \pi_\phi(a|s) - \Psi[Z(s,a)]] \tag{12}$$