Rebuttal for ErNESTO-gym: additional experiments

1 Different reward weighting with small clipping penalty

In this experiment, we tested with different reward weights, i.e., with $\{\theta_{trad} = 1, \theta_{op} = 1, \theta_{clip} = 0.1\}$ and $\gamma = 0.9$. The objective of the experiment is to evaluate our framework to demonstrate that RL algorithms can do better than trivial strategies. We also lowered the discount factor to foster a less farsighted approach to the problem.

For this experiment, we also introduce another algorithm, i.e., the soft-actor critic algorithm (SAC).

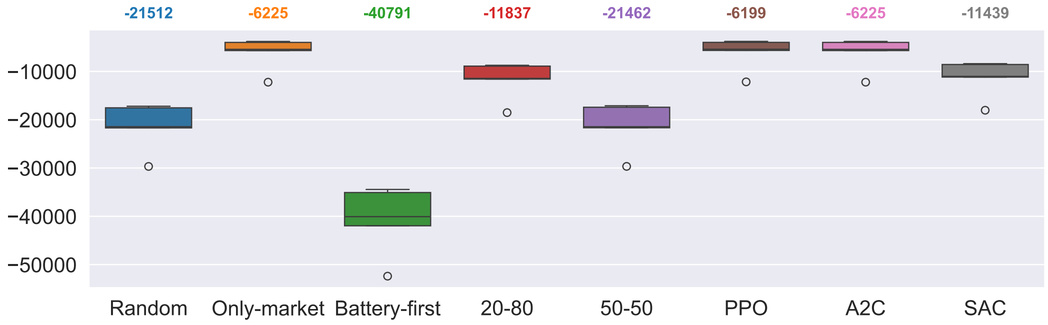

Figure 1: Boxplot representing the average return of each strategy. The values above each boxplot represent the mean of the relative algorithm across all the test experiments.

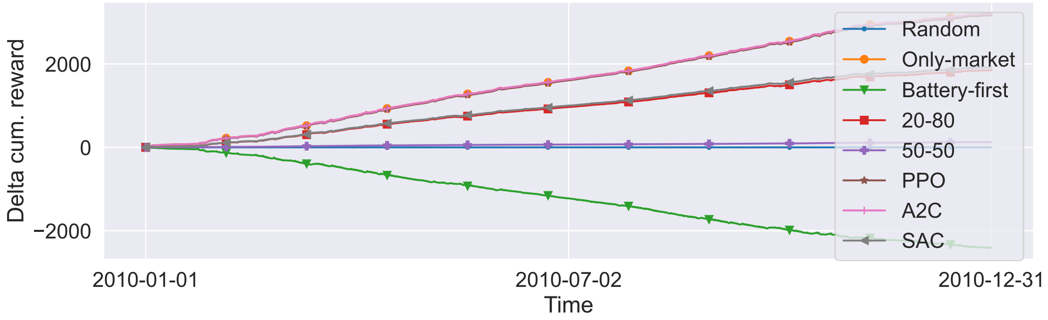

Figure 2: Average cumulative reward of each strategy compared with a baseline (*random actions*).

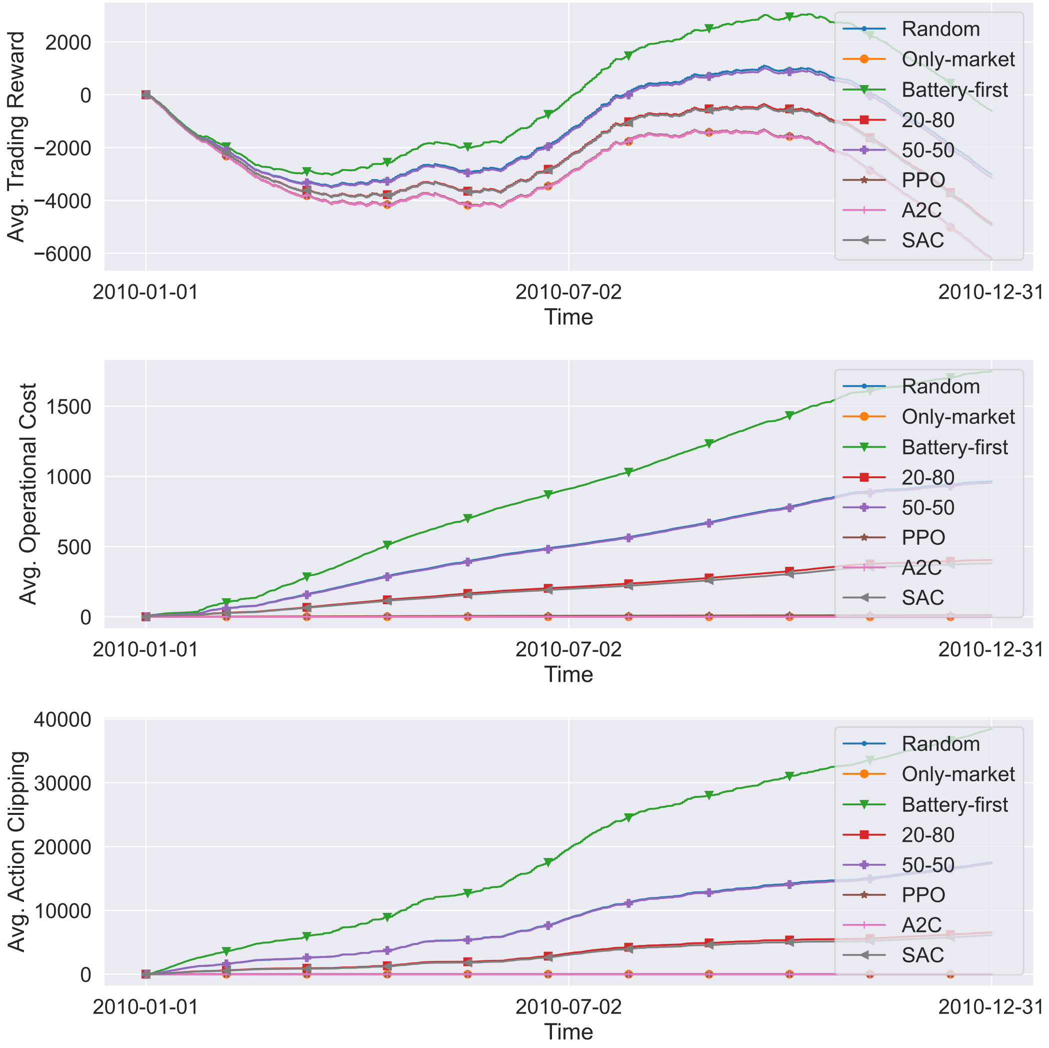

Figure 3: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm.

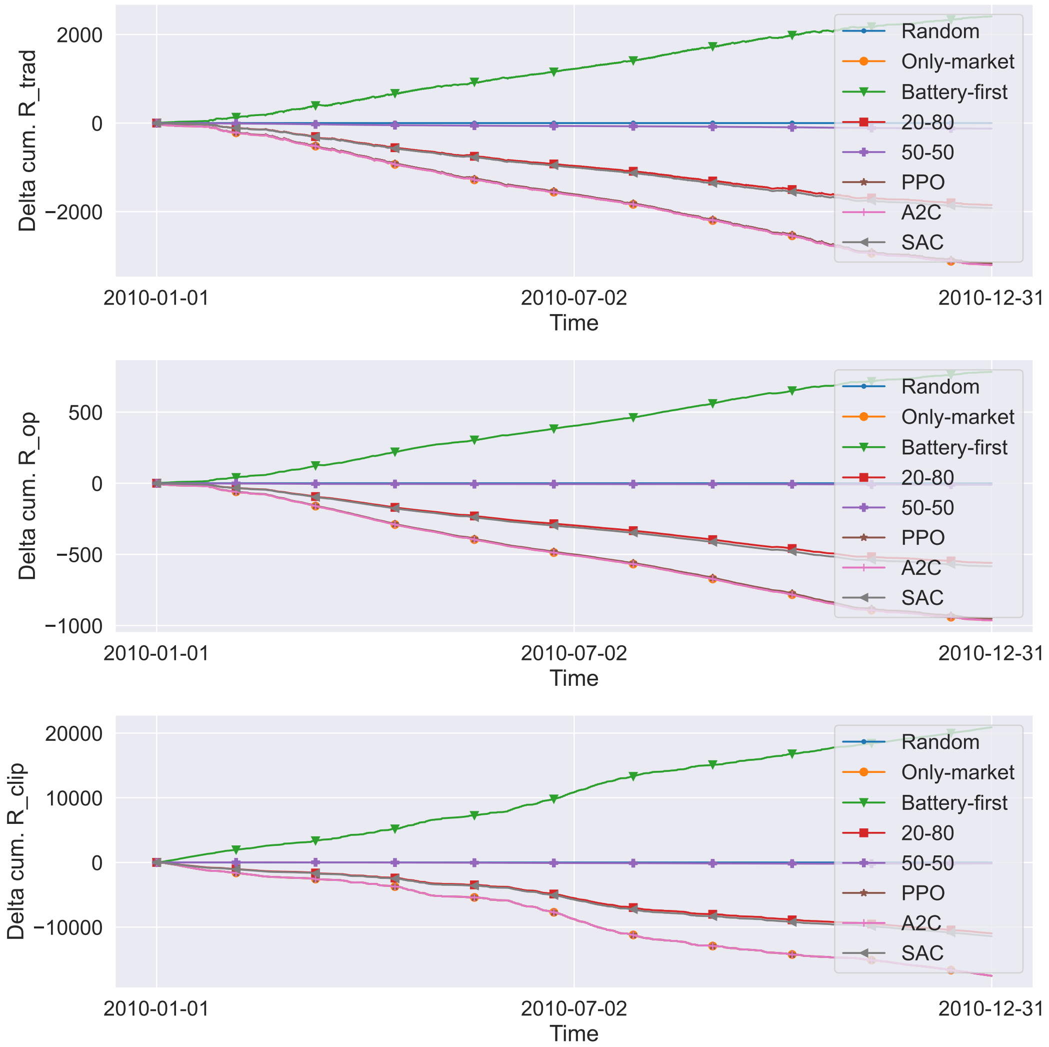

Figure 4: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm compared with a baseline (*random actions*).

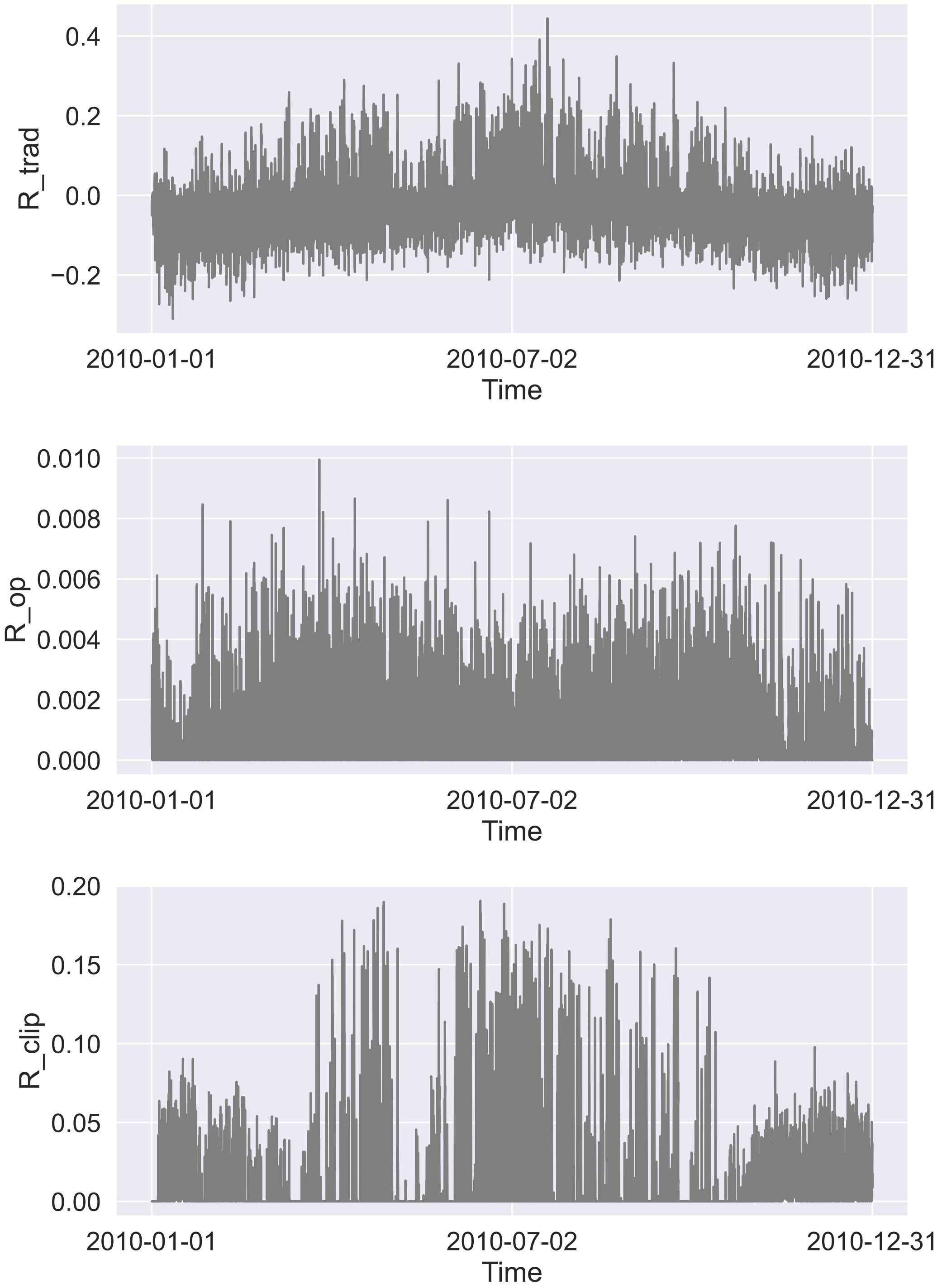

Figure 5: Analysis of SAC step rewards averaged among all the different evaluation scenarios.

2 Different reward weighting with no clipping penalty

In this experiment, as we did in the main paper, we try to evaluate the behavior of RL algorithms without considering the action clipping, which should prevent inappropriate usage of the battery. Here, we tested with the following reward weights, i.e., with $\{\theta_{trad} = 1, \theta_{op} = 1, \theta_{clip} = 0\}$ and $\gamma = 0.9$. It is interesting to notice the dissimilar behavior of PPO and A2C, which converge to different solutions to the problem. In this case, PPO performs better than A2C, which tends to choose unsafe actions.

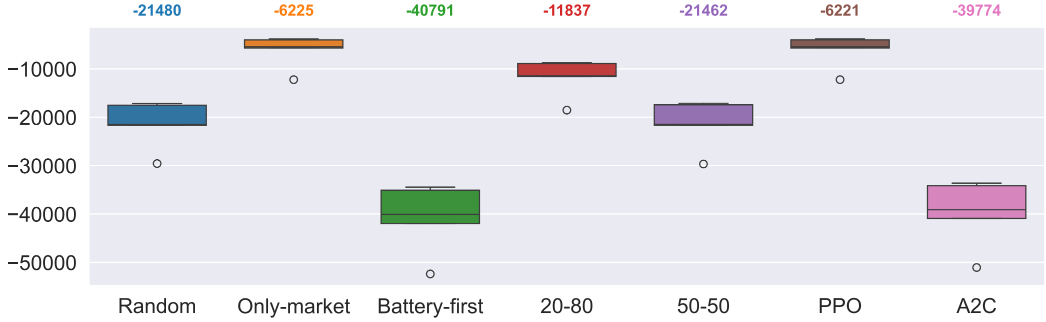

Figure 6: Boxplot representing the average return of each strategy. The values above each boxplot represent the mean of the relative algorithm across all the test experiments.

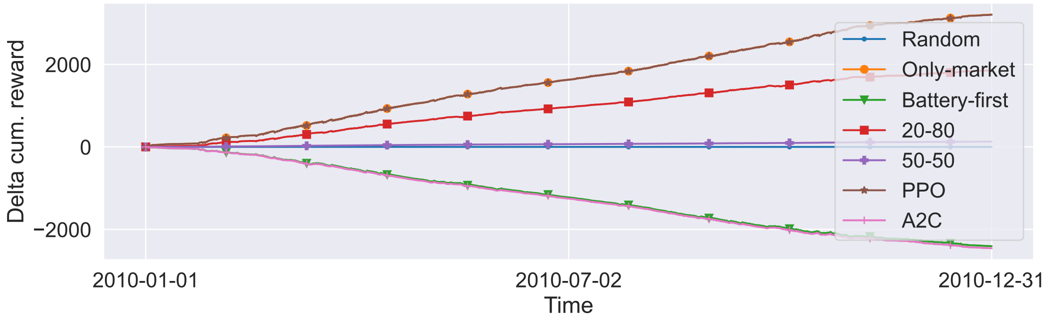

Figure 7: Average cumulative reward of each strategy compared with a baseline (*random actions*).

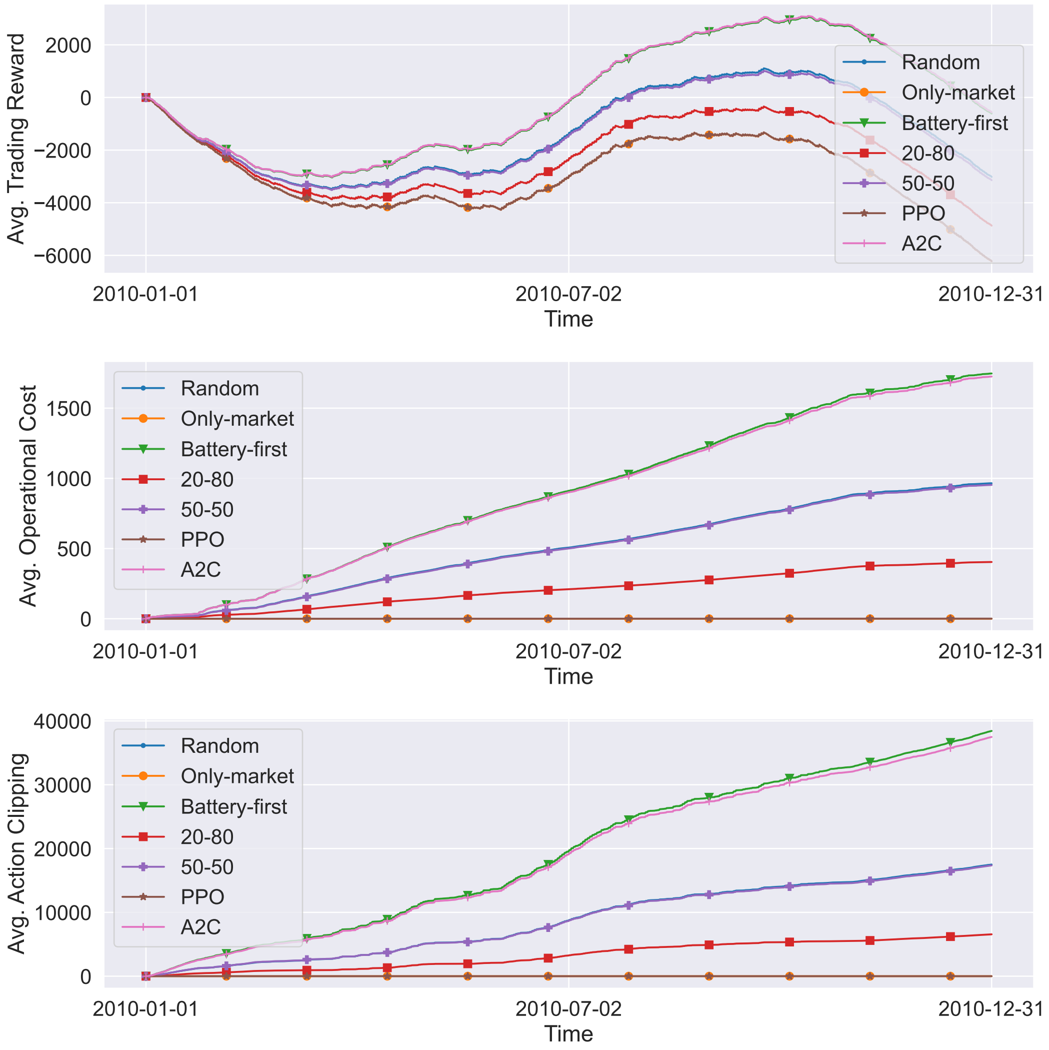

Figure 8: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm.

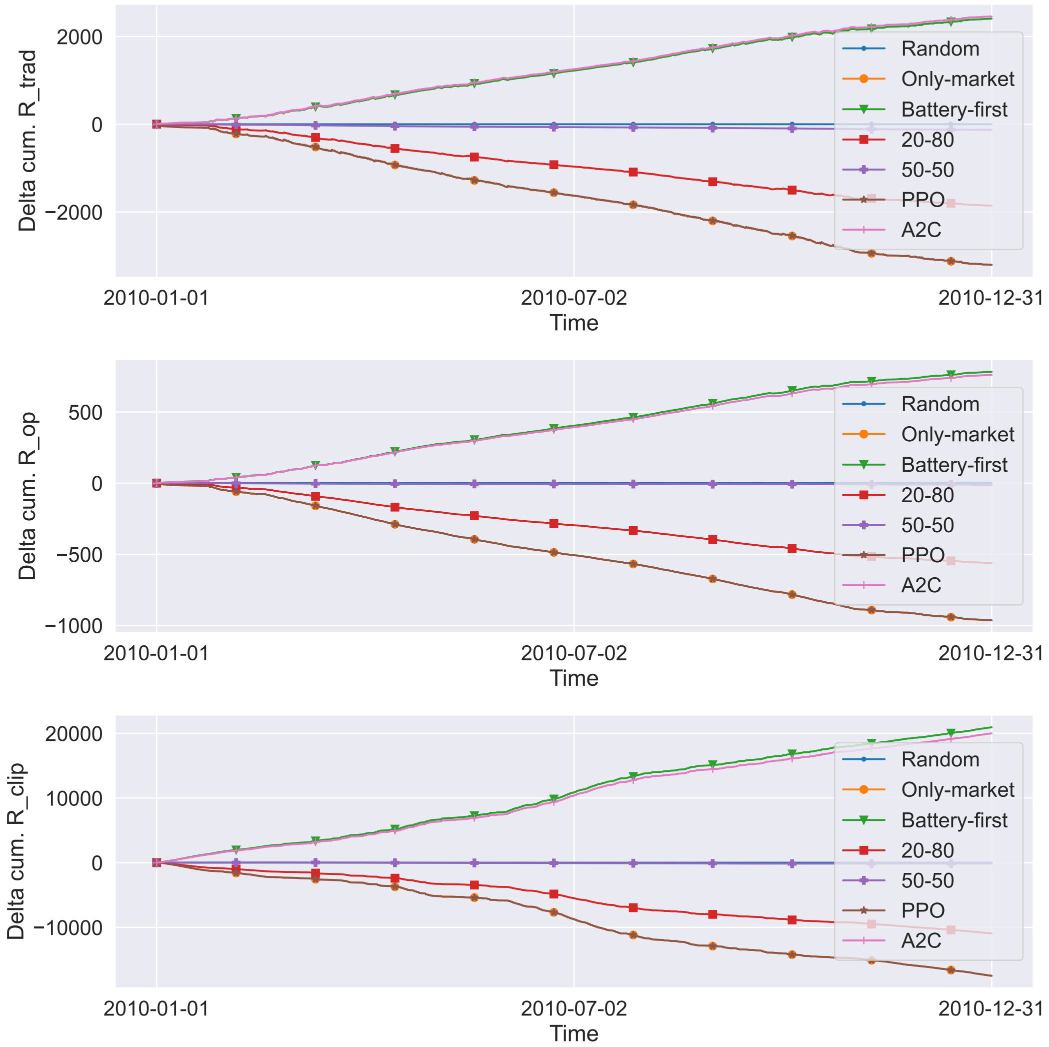

Figure 9: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm compared with a baseline (*random actions*).

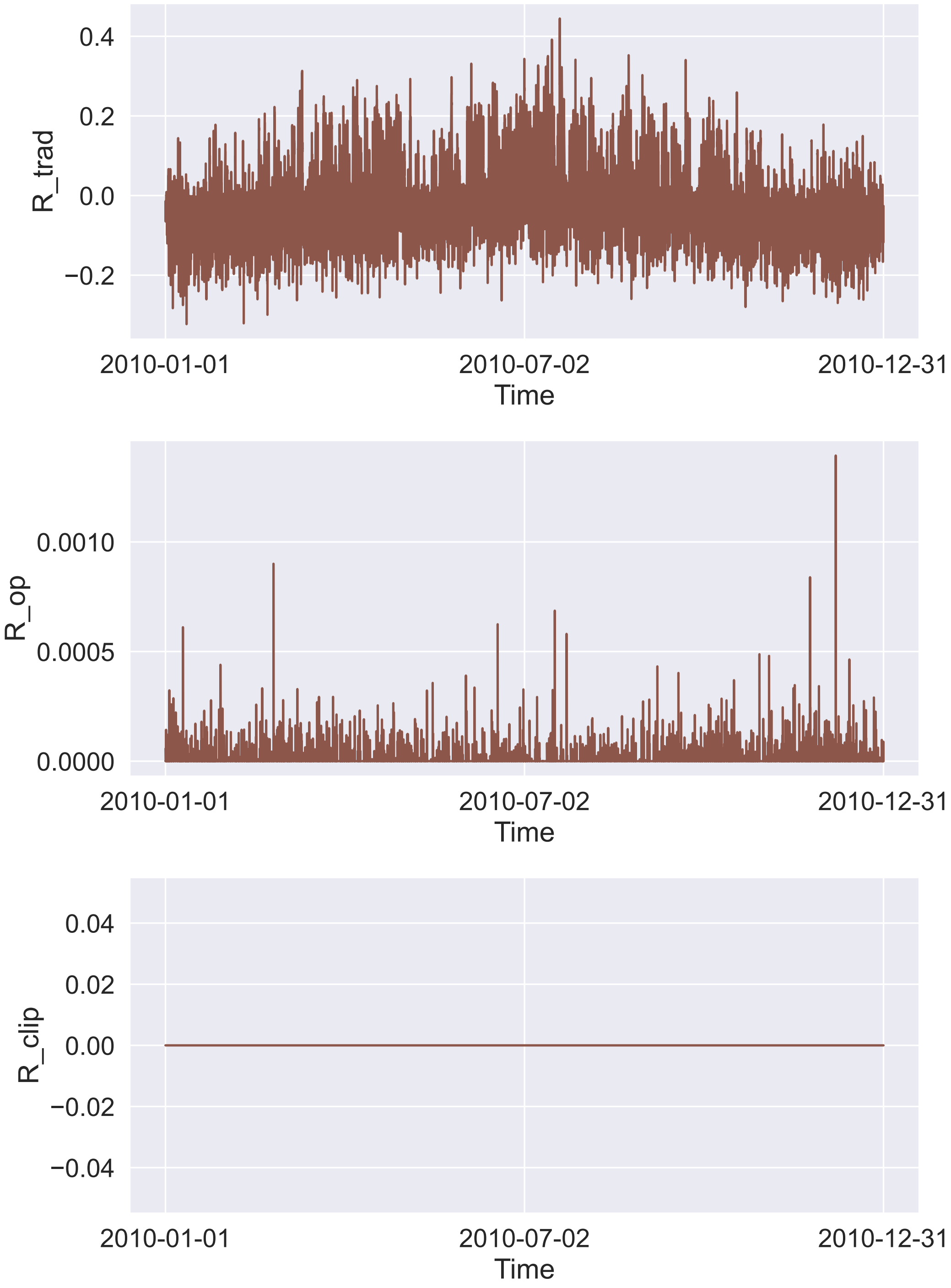

Figure 10: Analysis of SAC step rewards averaged among all the different evaluation scenarios.

3 Varying market conditions

To test with different market conditions, we trained and tested presented RL agents in a different scenario where the market price is multiplied by a factor of 5, increasing the gap between trading costs and revenues.

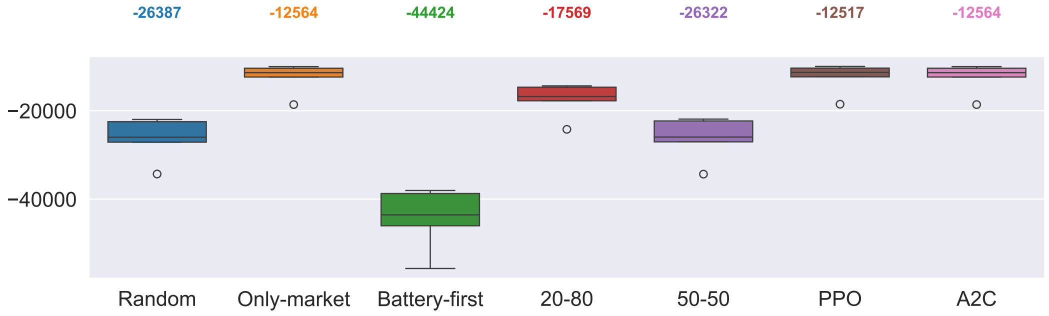

Figure 11: Boxplot representing the average return of each strategy. The values in the upper part represent the mean of the relative algorithm across all the test experiments.

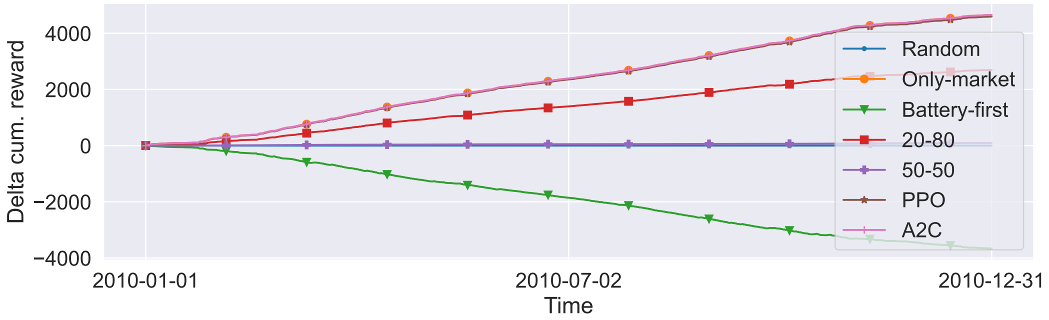

Figure 12: Average cumulative reward of each strategy compared with a baseline (*random actions*).

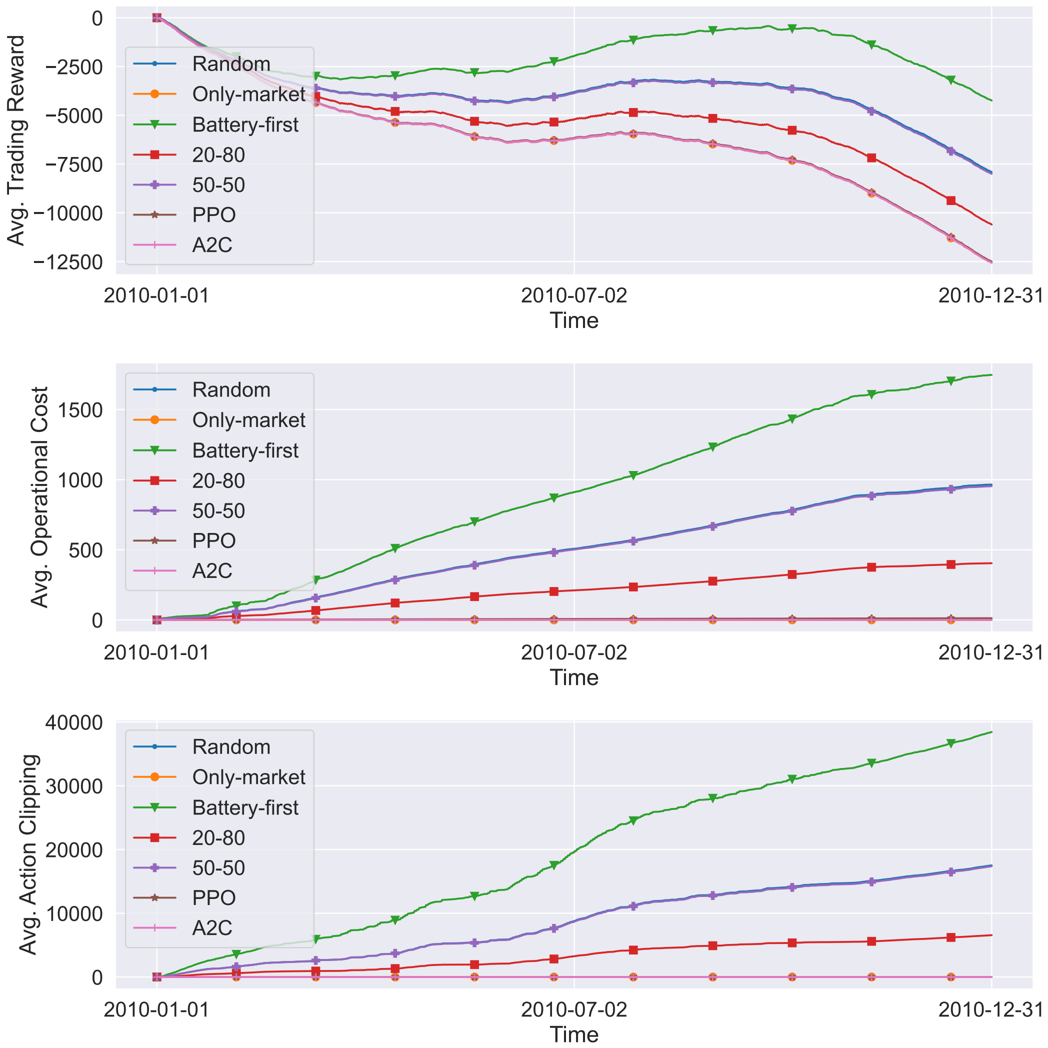

Figure 13: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm.

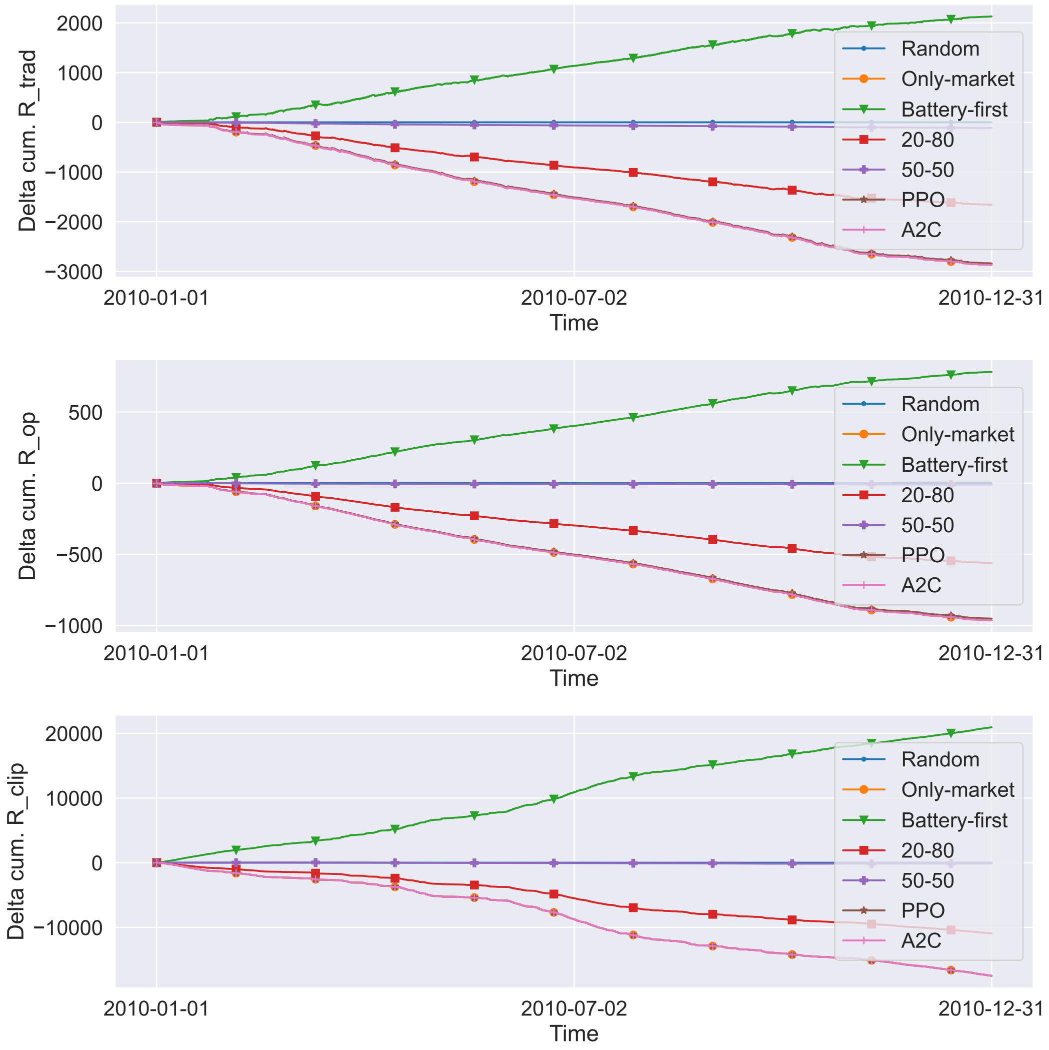

Figure 14: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm compared with a baseline (*random actions*).

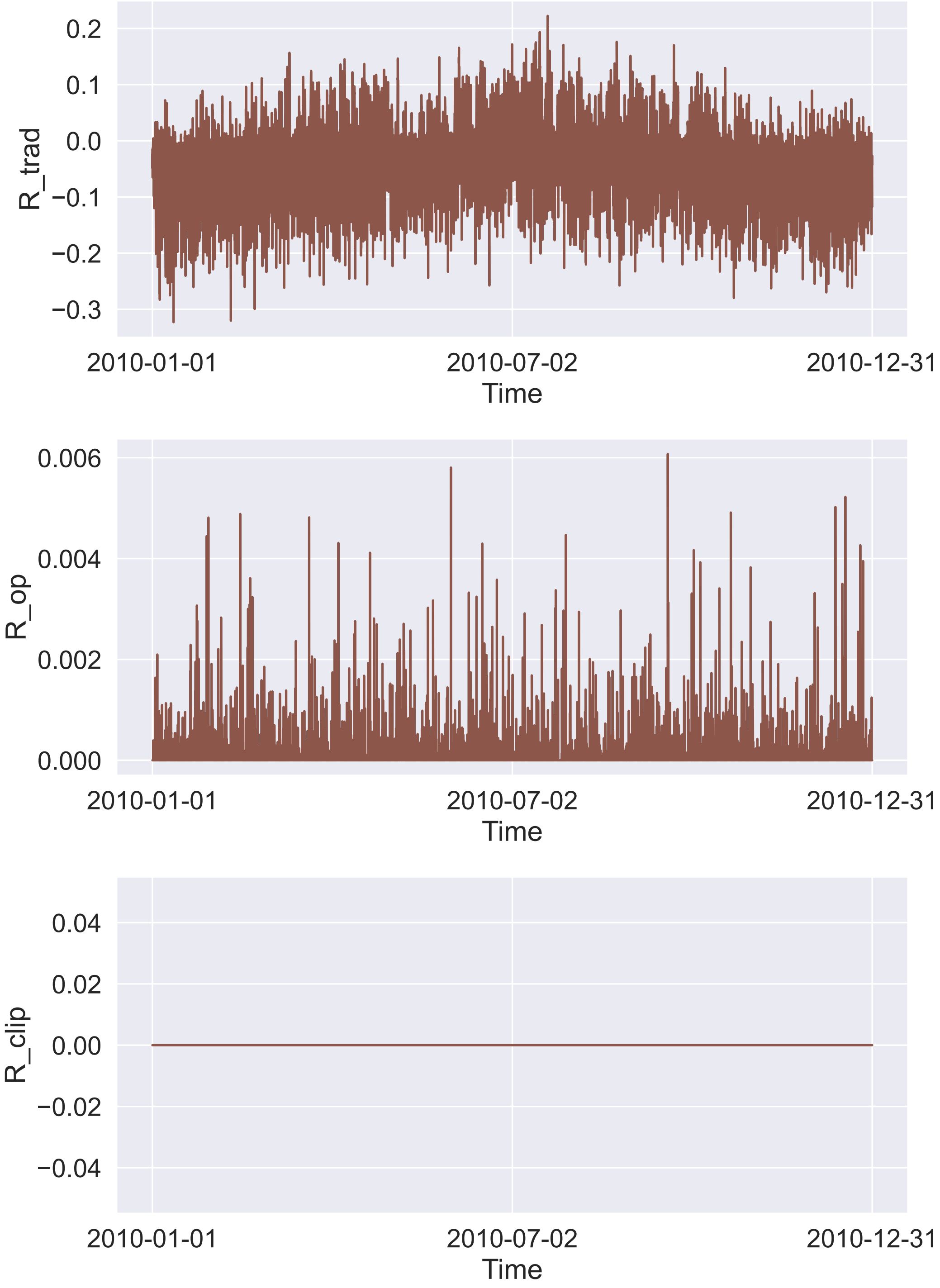

Figure 15: Cumulative average value of normalized reward terms (r_{trad}, r_{op} and r_{clip} from top to bottom) for each tested algorithm compared with a baseline (*random actions*).

4 Robustness analysis

To verify the robustness of the presented algorithms, we tested agents that had already been trained in different market conditions. In this scenario, the buying price from the energy market does not follow the trend of the real-world time series, but it is retrieved by sampling uniformly from a normal distribution $\mathcal{N}(\mu_{p^{\text{buy}}}, 2\sigma_{p^{\text{buy}}})$, where $\mu_{p^{\text{buy}}}$ is the mean of the real signal and $\sigma_{p^{\text{buy}}}$ its standard deviation.

Results show a good response from PPO and A2C to the modification within the scenario and show the robustness of these algorithms.

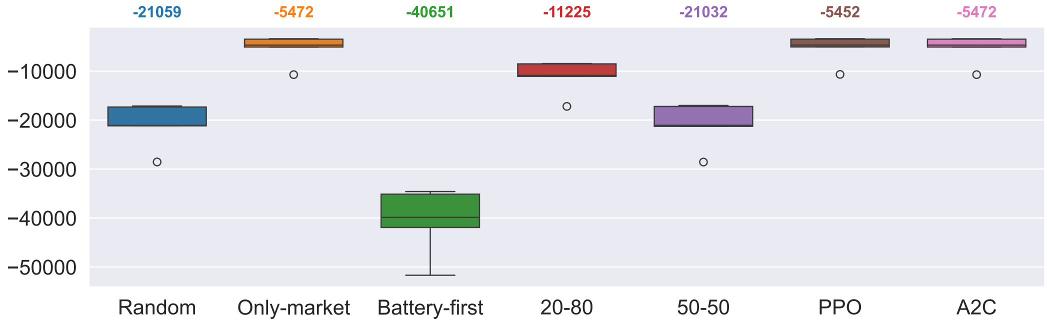

Figure 16: Boxplot representing the average return of each strategy. The values above each boxplot represent the mean of the relative algorithm across all the test experiments.

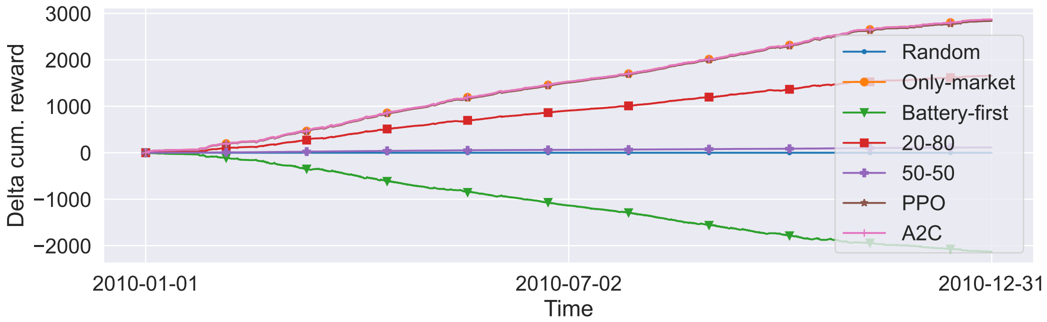

Figure 17: Average cumulative reward of each strategy compared with a baseline (*random actions*).

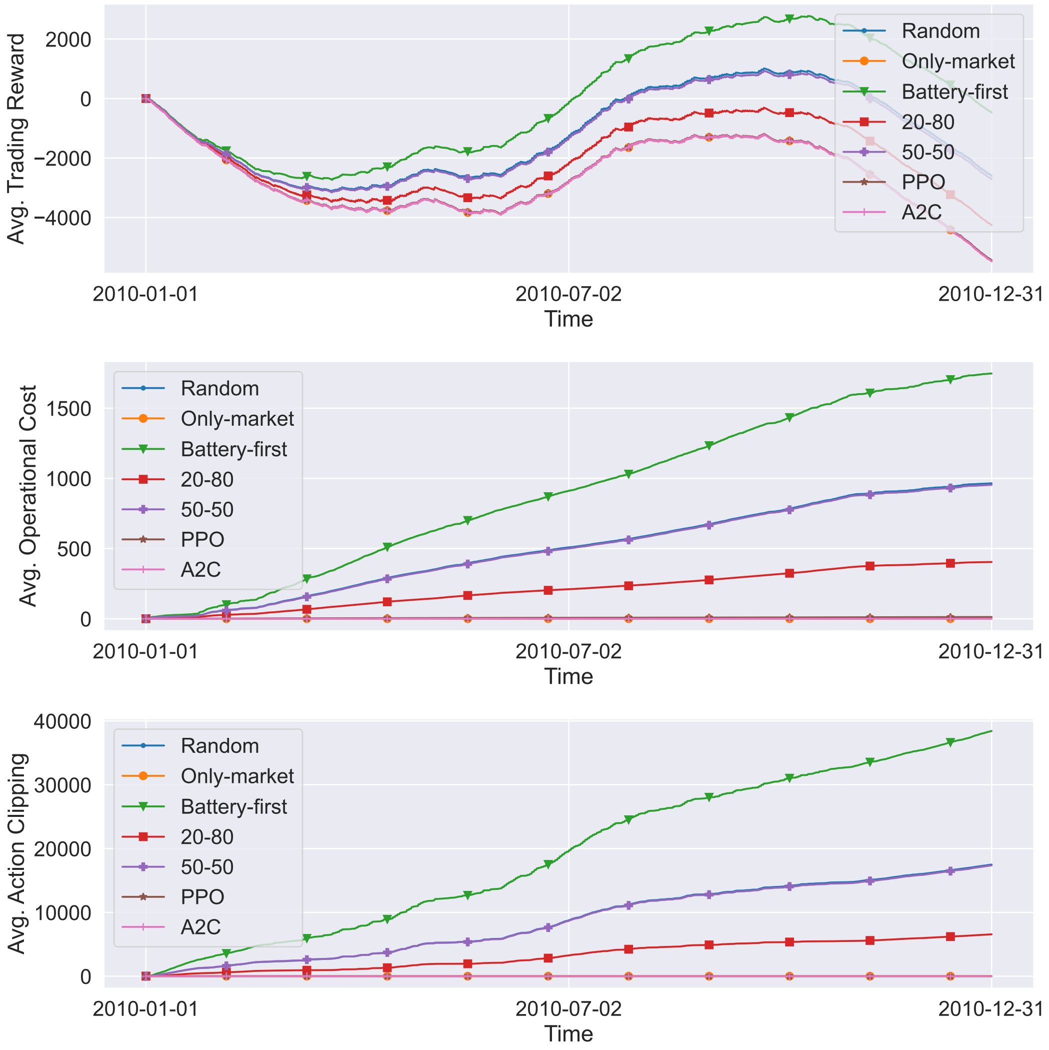

Figure 18: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm.

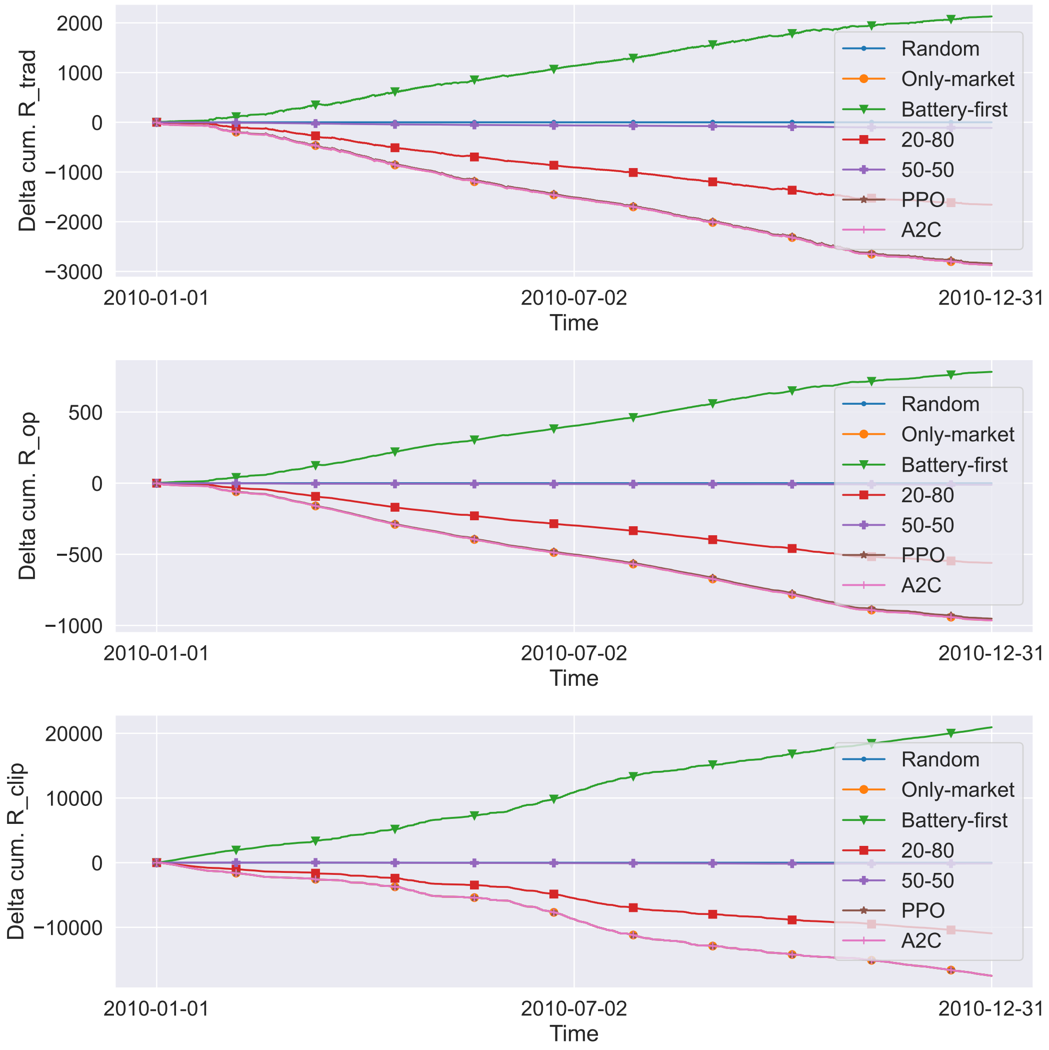

Figure 19: Cumulative average value of normalized reward terms (from top to bottom r_{trad}, r_{op}, and r_{clip}) for each tested algorithm compared with a baseline (*random actions*).

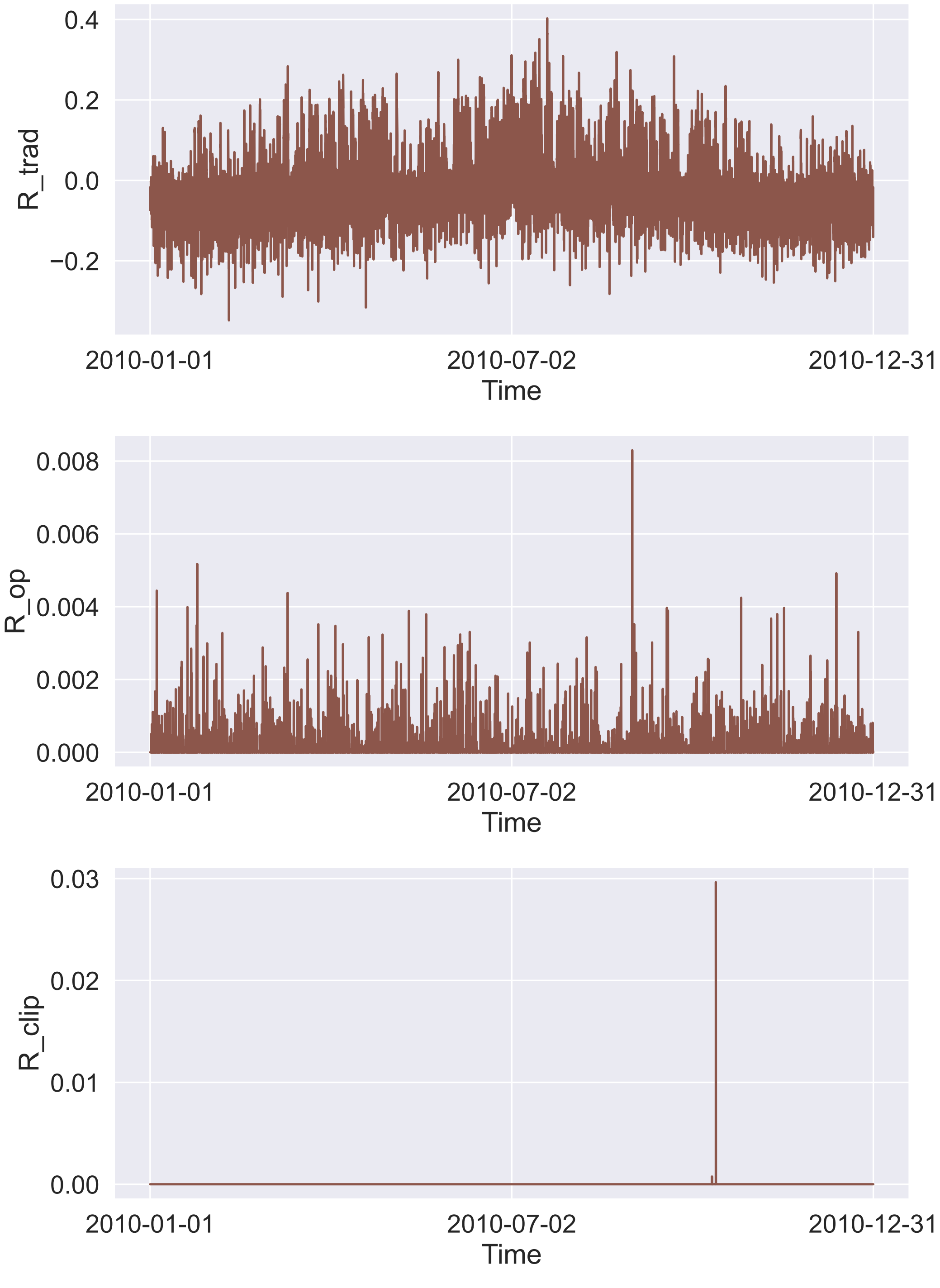

Figure 20: Analysis of PPO step rewards averaged among all the different evaluation scenarios.