Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
©2023 ESREL2023 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P488-cd

Rule-based deep reinforcement learning for optimal control of electrical batteries in an energy community

Roberto Rocchetta, Lorenzo Nespoli, Vasco Medici

University of Applied Sciences and Arts of Southern Switzerland, SUPSI, Mendrisio,
E-mail: roberto.rocchetta@supsi.ch; lorenzo.nespoli@supsi.ch; vasco.medici@supsi.ch

Saverio Basso, Marco Derboni, Matteo Salani

Dalle Molle Institute for Artificial Intelligence, IDSIA USI/SUPSI, Lugano, Switzerland,
E-mail: saverio.basso@idsia.ch; marco.derboni@idsia.ch; matteo.salani@idsia.ch

This work investigates rule-based controllers (RBCs) and reinforcement learning (RL) agents for managing distributed electrical batteries in a net-zero energy community (NZEC) and reducing costs and emissions for the community. The RBCs are based on deterministic rules, hence, may fail to adapt to new scenarios and uncertainties. On the other hand, RL agents learn from direct interaction with uncertain environments and can better adapt to new conditions. A novel RL approach is proposed, combining MaskPPO and a deep neural network, to avoid the exploration of unsafe/unprofitable actions and enhance control efficacy through accurate predictions of future demand. These new approaches are demonstrated on the *NeurIPS 2022 CityLearn challenge* where real-world data from a district in California are embedded within a simulator for distributed battery control. Points of strength and limitations of the different tools discussed. For comparison sake, an oracle-driven controller is also considered as it gives a reference best-achievable optimum for the challenge problem, i.e., lower bounds on costs and emissions reduction scores. Based on the results, RL agents generally offered robust control over the distributed batteries and often outperformed the rule-based controllers. Additionally, the combination of action masks and neural forecasters significantly improved the performance of the RL agents, bringing them very close to the scores achieved by the global optimum. A study of the model's robustness to seasonality changes concludes this work and further illustrates the generalization ability of controllers.

Keywords: Net-zero energy communities, Reinforcement Learning, Rule-Based control, Emissions, Peak shaving, Uncertainty

1. Introduction

Net-Zero Energy Communities are groups of buildings that, through intensive use of renewable sources and collaborative onsite management of available energy, minimize their dependency on external energy suppliers while enhancing the sustainability and profits of the community. For these groups of buildings, optimal control of the available storage units (ST) is essential to reduce demand peaks, fill demand valleys and, finally, lower consumptions, costs, and emissions Ullah et al. (2021). The problem of distributed ST control in NZEC has been addressed by various authors applying different techniques, e.g., deterministic rule-based control methods, Drgoňa et al. (2020), stochastic models, Medici et al. (2017), distributionally robust Gray et al. (2022) and risk-based methods Parvar and Nazaripouya (2022) and, recently, RL methods, Duque et al. (2022). RBCs define actions based on deterministic rules on known factors, e.g., time of day or load demand. Alternatively, rules can be informed by uncertain prediction of future quantities, such as loads or renewable PV productions Medici et al. (2017). RBCs are probably the most widely applied in practice due to their simple implementation. However, because of the uncertainty in future energy demand and renewable production, unforeseen scenarios can occur and undermine their effectiveness. If not optimized, RBC controllers usually lack robustness against uncertainties and do not adapt well to unforeseen operational conditions. Robustification strategies, that is, fine-parameter tuning and careful assessment of relevant uncertainties, are essential to guarantee a good performance. In contrast to RBCs, a reinforcement learning agent searches for an optimal control policy by directly interacting with an uncertain environment Sutton and Barto (2018). Because of this, RL agents are naturally more adaptable and better fitted to deal with natural variability and uncertainties. Advanced training algorithms, such as Proximal Policy Optimization (PPO), can be applied to handle mixed integer state-actions spaces and allows a district-centralized agent or building-specific controllers to be defined. Unfortunately, RL agents require long exploration periods and many observations before learning a useful control policy. Moreover, due to inherent safety concerns of applying exploratory action to real systems, high-fidelity simulators are required beforehand to train the agent offline pre-deployment Rocchetta et al. (2019). However, simulators are not always available. To overcome these limitations, algorithms like the Maskable PPO Garcıa and Fernández (2015) have been recently introduced and allow training the agent while only applying actions that are safe/feasible. Maskable PPO algorithms speed-up the learning process while reducing the risk of large economic losses and unsafe control trajectories, potentially leading to catastrophic failures.

The paper proposes novel rule-based and deep RL agents for managing distributed electrical storage in NZECs. The control objective is to minimize CO_2 emissions and electricity costs of an energy community while enhancing its ability to operate independently from the external power grid, e.g., by shifting and flattening the energy demand profile. Specifically, three optimized RBCs are proposed and prescribe deterministic control actions based on observed, predicted, and historical energy demands of the individual buildings. These three RBCs are compared to new deep RL agents trained with the Maskable PPO algorithm and equipped with action masks (constraints) and a neural predictor for future building energy demand. The action constraints ensure that only feasible actions are explored based on the observed State of Charge (SoC) and energy demand of batteries. In addition, a new mask function is also investigated and imitates the daily charge/discharge profiles from an optimal RBCs and showed promising results. The efficacy of the proposed controllers is demonstrated in the *NeurIPS 2022 - CityLearn Challenge*, a new gym environment for developing centralized and distributed RL controllers for NZEC. Furthermore, by equipping the agents with 10-hour-ahead energy demand predictions, additional improvements are observed in load-shaping reward, cost savings, and emission reduction. The proposed RBCs and RL agents are compared to a random controller and a no-storage case, and an Oracle-driven Model Predictive Controller (MPC) is introduced to find an approximation for the minimum achievable costs and emissions, i.e., a proxy for the global optimum. The paper concludes with an analysis of the controller's seasonal performance and its ability to handle new scenarios through generalization.

2. Preliminaries

This work considers sequential decision-making formulated as Partially Observable Markov Decision Processes (POMDPs),

$$(\mathcal{S},\mathcal{A},\mathcal{T},\mathcal{R},\Omega,\mathcal{O},f_0,T,\gamma),$$

where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an actions space, $\mathcal{T}$ is a set of conditional transition probabilities, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, Ω is a set of observations, $\mathcal{O}$, are conditional observation probabilities, T is the episode horizon, and γ is a reward discount factor. The goal of the decision-maker is to prescribe the best actions, i.e. an optimized control policy $\pi^\star$ so that the expected cumulative sum of discounted reward is maximised, $\mathbb{E}_{\pi^\star}\left[\sum_{t=0}^{T} \gamma^t r_t(a_t, o_t)\right]$.

Note that this problem admits both a centralized and decentralized version. In centralized control, a single agent makes decisions for the entire community based on a common policy, whereas in decentralized control, multiple agents make decisions based on local policies and information. Formally, a centralized battery control policy, π, maps observations to actions for the entire community ($\Omega \rightarrow \mathcal{A}$), while decentralized controllers, π_b, define steps for individual buildings ($\Omega_b \rightarrow \mathcal{A}_b$) based on local and shared observations. The optimized policy $\pi^\star$ maximizes the rewards for the district, whereas $\pi_b^\star$ focuses on individual buildings and may require coordination to optimize district-level rewards.

3. The CityLearn Gym environment

We adopt the simulation environment described by the *NeurIPS 2022 - CityLearn Challenge* Nweye et al. (2023). Five buildings in a set of buildings B define the NZEC, each equipped with batteries and PV generators. In an episode, the environment replays one year of historical data down-sampled to hourly resolution, i.e., time series of load demanded and PV produced by the building, weather, costs, and emission observations. For each step (an hour) in an episode, the energy balance for building b is computed and given by:

$$E_{h,b}^{NS} = E_{h,b} + E_{h,b}^{ST}(a), \ \forall b \in B \tag{1}$$

where $E_{h,b}^{NS}$ is the non-shiftable energy demand, $E_{h,b} = E_{h,b}^{M} - E_{h,b}^{PV}$ is the difference between energy demand and PV production (from data) and $E_{h,b}^{ST}(a)$ is the energy output of the battery, i.e., a function of control action a taken by the agent at the previous time step. Note that $E_{h,b}^{NS}$ must be satisfied by the external grid regardless of the actions taken and energy produced by the PVs.

3.1. *Observations and actions*

At each time step, the environment receives an action vector, $a = (a_1, .., a_b, ..., a_{n_b}) \in \mathcal{A}$ and returns an observation vector, $o = (o_1, ..., o_b, ..., o_{n_b})$, and a reward. Each battery/building receives an action $a_b \in [-1, +1]$, where $a_b > 0$ indicates charging and returns an observation, $o_b = (o_{sh}, o_{p,b}) \subseteq \mathbb{R}^{28}$ comprising a community-shared term, $o_{sh} = \left(t, W, \widehat{W}, \delta^{em}\right)$ and building-specific term, $o_p = \left(E^M, E^{PV}, SoC, E^{NS}, \delta^{el}, \hat{\delta}^{el}\right)$, i.e., private observations. The first observation vector includes a time-stamp, weather variables and perfect predictions, and carbon pressure δ^{em} that is an emission cost expressed in $[\frac{kg_{CO2}}{kWh}]$. The private term includes energy demand, PV production, SoC, electricity price and three perfect forecasts for the next 24 h prices. Note the high-dimensionality of the problem with $o \in \mathbb{R}^{140}$. However, it can be lowered to $\mathbb{R}^{44}$ by removing shared observations and identical electricity price duplicates for the 5 buildings.

3.2. *Reward function*

The following modified reward function is used in this work:

$$r_h = -\sum_{b=1}^{n_b} \left(C_{b,h}^{el} + C_{b,h}^{em} + r_{b,h}^{sh}\right) \tag{2}$$

where C^{el} is the clipped cost of energy demand, C^{em} is the clipped cost of carbon emissions, and $r_{b,h}^{sh} = \frac{E_{b,h}^2 - (E_{b,h}^{NS})^2}{\beta}$ is a load shaping reward, a penalizing term for missed peak-shaving and valley-filling. The parameter β defines asymmetric rewarding of peak-shaving, i.e., $\beta = 10$ if $E_{b,h}^{NS} > 0$, and valley filling, $\beta = 100$ if the net demand is negative.

3.3. *Key performance indicators*

We use normalized KPIs to evaluate the controllers with respect to a reference policy π_0, that is, a case without batteries. We adopt the cost metric $m_1(\tau;\pi) = \sum_{h=0}^{\tau} C_{b,h}^{el}(\pi) \ / \ \sum_{h=0}^{\tau} C_{b,h}^{el}(\pi_0)$, defining the average normalized electricity price until step τ

and, similarly, we also adopt a metric m_2 and m_3 which are as normalized emission cost and a normalized grid cost, respectively. The latter, relates to congestion management and stability issues and is a function of ramping and load factors, see the original challenge description for further details.

4. The proposed approach

The proposed load predictor, rule-based controllers, oracle-driven MPC, and deep RL agents are introduced next.

4.1. *Electric load forecasting*

We predict the energy demand for each $b \in B$ in the next H hours. For this, we train a neural network model on the observations collected from the last K steps. The resulting NN predictor is given by:

$$\left(\hat{E}_{h+1}, ..., \hat{E}_{h+H}\right) = f\left(o_h, ..., o_{h-K}; \omega\right), \quad (3)$$

where ω are trainable parameters, and $\hat{E}_{h+j}$ are 5-dimensional vectors of load predictions for j-hours ahead (one for each building). As an example for $H = 1$ and $K = 12$, a fully connected network with 312 input nodes (plus flattening and batch normalization layers), three fully connected hidden layers with (600, 300, and 80 nodes) and five output nodes, for a total of 393,209 trainable parameters achieved very high prediction performance. Parameter tuning, pruning, and architectural optimization can further improve accuracy and efficiency but are out of the scope of this work and not further considered.

4.2. *Optimized Rule-based controllers*

The optimization problem for the RBC is defined as follows:

$$\min_{p} \sum_{i=1}^{3} m_i(\tau; \pi(o; p)) \quad (4)$$

where p is a vector of parameters defining the policy $\pi(o; p)$ and the objective is to minimise the sum of normalized electricity cost, carbon emissions, and grid stability-stability-related scores. Three RBC policies are considered: (i) **Persist**, where actions are based on the present energy demand, (ii) **Predict**, where actions are selected based on the next-hour prediction $\hat{E}_{h+1}$, and (iii) μ-**daily** that select actions based on the hour of the day. The mathematical definition of an action a_b in (i) and (ii) is given by:

$$a_b = \begin{cases} p_1 E_b & \text{if } E_b \le 0 \\ p_2 E_b & \text{if } E_b > 0 \end{cases}, \ \forall b \in B, \quad (5)$$

where E_b refers to the observed net demand at time h for the Persist agent and to the predicted demand at $h + 1$ for the Predict policy. The μ-daily defines an optimal average daily charging/discharging profile as follows:

$$a_b = \begin{cases} p_{1,b} & \text{if } h = 1 \\ ... & ... \\ p_{24,b} & \text{if } h = 24 \end{cases}, \forall b \in B, \quad (6)$$

where $24 \times n_b$ parameters must be optimized and the μ-daily actions only depend on the hour of the day.

4.3. *Oracle-driven MPC*

We propose an oracle-driven MPC, a linear bi-objective optimization model, where future net energy is known, that minimizes both the electricity (z_1) and emission costs (z_2) in a lexicographic order over the set of buildings B and the whole time horizon T. The optimal solution represents a lower bound with respect to the minimum electricity costs and emissions achievable for this challenge. In our formulation, we describe our main control actions with variables $c_{b,h}$ and $d_{b,h}$ which represent, respectively, the energy charged and discharged by the batteries in each building $b \in B$ and for each time step $h \in T$. The first objective function z_1 is defined as follows:

$$z_1 = \min \sum_{h \in T} C_h^{el} \quad (7)$$

$$C_h^{el} \ge \delta_h^{el} \sum_{b \in B} (E_{h,b} + c_{b,h} - d_{b,h}) \qquad \forall h \in T$$

where C_h^{el} is a non-negative community electrical cost and δ_h^{el} is the electricity price at time h. The second objective function (z_2) is instead the following:

$$z_2 = \min \sum_{b \in B, h \in T} C_{b,h}^{em} \tag{8}$$

$$C_{b,h}^{em} \geq \delta_h^{em} \left(E_{h,b} + c_{b,h} - d_{b,h}\right) \quad \forall b \in B, h \in T$$

where $C_{b,h}^{em}$ is the non-negative emission cost for a given $b \in B$ and $h \in T$, and δ_h^{em} is the carbon intensity at step h. Additional constraints are imposed on the energy produced/demanded by the batteries and on the SoC at each time set. These constraints are given by:

$$\begin{aligned} &0 \leq c_{b,h} \leq c_b^{max} && \forall b \in B, h \in T \\ &0 \leq d_{b,h} \leq d_b^{max} && \forall b \in B, h \in T \\ &SoC_{h,b} = SoC_{h-1,b} + \left(\zeta_b c_{b,h} - \frac{1}{\zeta_b} d_{b,h}\right) && \forall b \in B, h \in T \setminus \{1\} \\ &0 \leq SoC_{h,b} \leq q_b && \forall b \in B, h \in T \end{aligned}$$

where c_b^{max} and d_b^{max} are charging and discharging maximum energy for batteries, $SoC_{1,b}$ is the initial state-of-charge, q_b is the capacity of the battery, and ζ_b is the battery efficiency. In our linear formulation, the efficiency ζ_b is an upper bound to the one employed by the CityLearn environment. Therefore, our solutions are lower bounds to the optima achievable in the challenge.

4.4. *The rule-based deep RL agents*

In this work, we adopt the popular PPO actor-critic method proposed by Schulman et al. (2017) and extend it with safe action masks by adopting the MaskablePPO algorithms recently introduced by Huang and Ontañón (2022). During all the phases of the analysis, we used *Stable-Baselines3* package to train the agents, and always apply a learning rate $\alpha = 0.001$, a discount factor $\gamma = 0.99$, and a truncate the episode at 4000 steps (hours).

4.4.1. *Maskable PPO*

Three constraint functions, that avoid unwarranted battery actions, are defined as follows:

Mask-1: $a_b \in [-SoC_b, 1 - SoC_b] \subset \mathcal{A}_b$;
Mask-2: $a_b \in [0, 1 - SoC_b]$ if $E_b \leq 0$ else $a_b \in [-SoC_b, 0]$;
Mask-3: $a_b \in [p_{b,h} \pm \epsilon]$ for $h = 1, ..., 24$, where ϵ defines a half-width interval around $p_{b,h}$.
Mask-1 is based on the battery SoC, Mask-2 combines SoC and energy demand, whilst Mask-3 constraints the exploration in the proximity of the optimized μ-daily policy. We expect Mask-3 to inherit good performance of the μ-daily, but also be more robust and generalize better to new scenarios.

4.4.2. *Maskable PPO with predictor*

In addition to the three mask functions defined in the previous section, we attempt to further enhance the agent performance by combining them with the energy predictor introduced in section 4.1. In our approach, energy demand predictions for the next 10 hours, $\left(\hat{E}_{h+1}, ..., \hat{E}_{h+10}\right)$, are combined with a subset of 28 factors from o, for a total of 78 observations.

5. Results and discussions

The data set of the NeurIPS 2022 CityLearn Challenge is used to train the predictor, see section 4.1, to optimize the three RBCs, 4.2, the Oracle-driven MPC, 4.3, and to train the RL agents introduced in section 4.4.

5.1. *Comparison of the control policies*

Table 1 shows the results for the optimized RBC and RL agents and the Opracle-driven MPC and compares the action spaces, optimizers, cost reduction m_1, emission reduction m_2, load shaping reward and other KPIs. For comparison sake, we also show the scores of a reference policy without batteries (NoST), a random policy (Rnd), and two expert-based charge-at-night (CaN) and discharge-at-night (DaN) controllers.
Random and experience-based agents perform very poorly, even worse than a case without batteries in terms of costs $m_1 > 1$, emissions $m_2 > 1$, grid-related ramping and load factors $m_3 > 1$. This result supports the

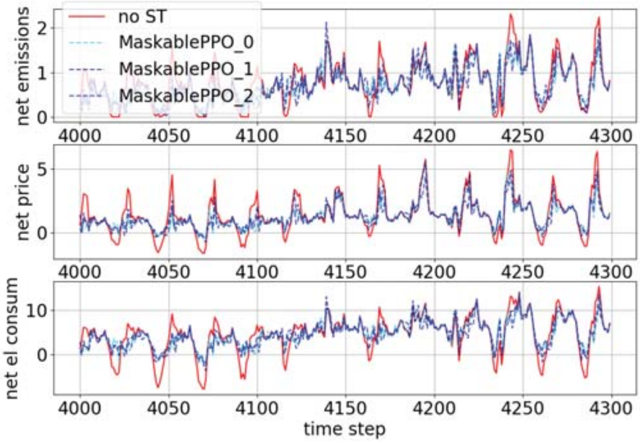

Fig. 1. Flattening of the emissions, costs, and energy consumption profiles achieved by the Maskable PPO agents equipped with the load predictor (blue dashed lines) versus the original case without batteries (the no-storage case in red). Also, around step 4150, note the relatively low performance due to consecutive days with a low PV production.

Table 1. Summary of the control policies including, names, state space (discrete or box continuous), learning model, and resulting key performance indicators, e.g., $\mu_{R_{sh}} = \mathbb{E}[\sum_b r^{sh}_{h,b}]$ and $\mu_{R_{tot}} = \mathbb{E}[r_h]$. The RL agents combined with the neural predictor use a 10h-ahead net load predictor and a subset of shared and private observations. The optimized p for Predict and Persist are $(-0.120, -0.155)$ and $(-0.107, -0.142)$, respectively.

Name	Type	Optimizer	$\mathcal{A}_b$	m_1	m_2	m_3	$\mu_{R_{sh}}$	$\mu_{R_{tot}}$	$\frac{\sum_i m_i}{3}$
NoST	Baseline	-	Dis(1)	1.0	1.0	1.0	0	-1.530	1.0
Rnd	Random	-	Box(-1,1)	1.34	1.93	2.57	-2.206	-5.086	1.95
CaN	RBC	Expert-based	Dis(2)	1.069	1.179	1.075	-0.190	-1.894	1.107
DaN	RBC	Expert-based	Dis(2)	1.135	1.058	1.068	-0.079	-1.793	1.087
Persist	RBC	Annealing	Dis(2)	0.788	0.854	0.992	0.1304	-1.148	0.878
Predict	RBC	Annealing	Dis(2)	0.804	0.888	0.981	0.1093	-1.216	0.891
μ-daily	RBC	Annealing	Dis(24)	0.695	0.976	0.947	0.0102	-1.369	0.873
PPO	RL	PPO	Box(-1,1)	1.001	1.001	1.001	-0.0	-1.53	1.001
Mask-1	RL	MaskPPO	Dis(51)	0.753	0.918	0.983	0.041	-1.229	0.885
Mask-2	RL	MaskPPO	Dis(51)	0.780	0.924	0.970	0.0324	-1.235	0.891
Mask-3	RL	MaskPPO	Dis(51)	0.736	0.921	0.976	0.0903	-1.233	0.877
PPO-pred	RL	PPO	Box(-1,1)	0.930	1.005	0.975	0.019	-1.510	0.970
Mask-1-pred	RL	MaskPPO	Dis(51)	0.705	0.866	0.912	0.142	-0.846	0.828
Mask-2-pred	RL	MaskPPO	Dis(51)	0.720	0.870	0.931	0.134	-0.870	0.841
Mask-3-pred	RL	MaskPPO	Dis(51)	0.694	0.874	0.927	0.128	-0.843	0.832
Oracle-Cost	MPC	Linear Prog	Box(-1,1)	**0.542**	0.752	0.979	-	-	0.758
Oracle-Emis	MPC	Linear Prog	Box(-1,1)	0.625	**0.695**	1.137	-	-	0.819

need for optimized battery control strategies. In the last two rows of 1, we present the results of the Oracle-driven MPC, which define a lower bound on $m_1 \geq 0.542$ (Oracle-Cost) and a lower bound on $m_2 \geq 0.695$ (Oracle-Emis). These are the best achievable costs and emissions in the challenge and will provide a useful reference for future comparisons.

Our optimized RBC and RL agents achieved an overall good performance, with better KPIs compared to the NoST baseline. The optimized Predict and Persist agents perform well, especially the latter that has the highest load shaping reward (0.13) and lowest emissions (0.854) among the other optimized. On the other hand, the μ-daily policy aggressively tries to minimize electricity costs but with a substantially lower performance in terms of load-shaping reward (0.01) and emission score (0.976). Nevertheless, thanks to an electricity bill reduction of more than 30 % (0.695), the μ-daily agent achieves the highest mean score (0.873) among the three optimized RBCs.

The RL agents without predictor performed quite well compared to the baseline, however, their performance is slightly worse than the optimized RBCs, with similar mean m_i scores but for lower load-shaping reward (0.04-0.09). This was probably due to the early truncation of the learning due to computational time constraints. Interestingly, a simple PPO agent very often converged to a bad sub-optimal policy equivalent to a case without batteries case. These RL agents, when equipped with a predictor, led to the best control performances. For instance, Masl-1-pred achieve the highest load-shaping reward (0.142), lowest grid cost (m_3) and lowest average KPI score (0.828). It is also interesting to look at the results of the Mask-3-pred agent, which constrains the policy search informed by the μ-daily. Mask-3-pred inherited the same good cost reduction score (0.694) for the rule-based policy, however, it also performs much better in all the other metrics, hence providing a more robust control compared to the deterministic RBC.

5.2. *Generalization and seasonality effects*

The generalization ability of the three MaskPPO agents equipped with the load predictor is compared based on their ability to predict out-of samples data from unseen seasons. A validation set is defined with data from half of January until July, i.e., for h >4000 and relative rolling variance reduction, $1-Var_{\pi^\star}(E^{NS})/Var_{\pi_0}(E^{NS})$, is presented and discussed. Note a significant reduction in the rolling variance throughout the year, which indicates good generalization of the RL agents. A higher reduction is observed during spring, autumn, and summer (50-80 %) and lower during winter (30-60 %). This seasonality trend was expected due to the lower PV production in winter. Nonetheless, during sunny days in winter, the variance can reduce up to high as 80 %. The three agents perform very similarly during different seasons, with a slightly better performance of the MPPO3 during March and April and a better performance of the MPPO1 during winter (November to February)

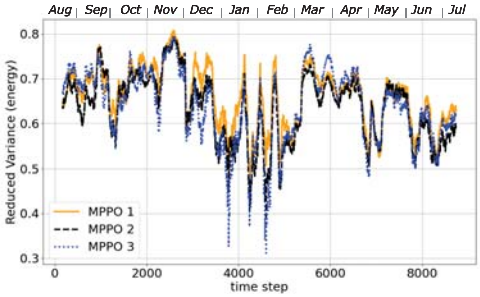

Fig. 2. Reduction of the weekly variability of the non-shiftable energy demand (district level) achieved by the three Maskable PPO agents equipped with the load predictor.

6. Conclusion and future directions

This paper proposed RBC and RL agents for optimizing electric batteries in net-zero energy communities. Tested on the CityLearn simulator (NeurIPS-2022), the results show that improper use of storage units can lead to higher costs and emissions possibly damaging the community. An oracle-driven MPC defines the lowest cost (0.542) and emission reduction (0.695) achievable ad the RL agents perform closest to the global optimum. The MaskPPO combining action masks and a neural forecaster reduces electricity costs by 30% (0.694) and emissions by 12.6% (0.874). They also improve grid stability and load profile shape. Future research aims to test the controllers on more complex energy communities and explore centralized and multi-agent RL with model-predictive controllers and probabilistic load forecasters. Lower KPIs may be achieved with higher-resolution sampling, e.g., a 1-minute frequency, and by applying multi-agent RL techniques. This will be part of future extensions.

Acknowledgement

This research has been partially supported by the Swiss Federal Office of Energy with the project SI/502557 ASTRO-PROCOM – "*Optimised battery management to maximise lifetime by pooling stored energy in a community of prosumers*" and partly by the Swiss National Science Foundation under NCCR Automation (grant agreement 51NF40 180545)

References

Drgoňa, J., J. Arroyo, I. Cupeiro Figueroa, D. Blum, K. Arendt, D. Kim, E. P. Ollé, J. Oravec, M. Wetter, D. L. Vrabie, and L. Helsen (2020). All you need to know about model predictive control for buildings. *Annual Reviews in Control 50*, 190–232.

Duque, E. M. S., J. S. Giraldo, P. P. Vergara, P. Nguyen, A. van der Molen, and H. Slootweg (2022). Community energy storage operation via reinforcement learning with eligibility traces. *Electric Power Systems Research 212*, 108515.

Garcıa, J. and F. Fernández (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research 16*(1), 1437–1480.

Gray, A., A. Wimbush, M. de Angelis, P. Hristov, D. Calleja, E. Miralles-Dolz, and R. Rocchetta (2022). From inference to design: A comprehensive framework for uncertainty quantification in engineering with limited information. *Mechanical Systems and Signal Processing 165*, 108210.

Huang, S. and S. Ontañón (2022). A closer look at invalid action masking in policy gradient algorithms. *The International FLAIRS Conference Proceedings 35*.

Medici, V., M. Salani, L. Nespoli, A. Giusti, M. Derboni, N. Vermes, A. E. Rizzoli, and D. Rivola (2017). Evaluation of the potential of electric storage using decentralized demand side management algorithms. *Energy Procedia 135*, 203–209.

Nweye, K., S. Sankaranarayanan, and Z. Nagy (2023). MERLIN: Multi-agent offline and transfer learning for occupant-centric energy flexible operation of grid-interactive communities using smart meter data and citylearn. *arXiv*.

Parvar, S. S. and H. Nazaripouya (2022). Optimal operation of battery energy storage under uncertainty using data-driven distributionally robust optimization. *Electric Power Systems Research 211*, 108180.

Rocchetta, R., L. Bellani, M. Compare, E. Zio, and E. Patelli (2019). A reinforcement learning framework for optimal operation and maintenance of power grids. *Applied Energy 241*, 291–301.

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

Sutton, R. S. and A. G. Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Ullah,

K., V. Prodanovic, G. Pignatta, A. Deletic, and M. Santamouris (2021). Technological advancements towards the net-zero energy communities: A review on 23 case studies around the globe. *Solar Energy 224*, 1107–1126.