Generalising Battery Control in Net-Zero Buildings via Personalised Federated RL

Nicolas Cuadrado [*1] **Samuel Horváth** [1] **Martin Takáč** [1]

Abstract

This work studies the challenge of optimal energy management in building-based microgrids through a collaborative and privacy-preserving framework. We evaluated two common RL algorithms (PPO and TRPO) in different collaborative setups to manage distributed energy resources (DERs) efficiently. Using a customized version of the CityLearn environment and synthetically generated data, we simulate and design net-zero energy scenarios for microgrids composed of multiple buildings. Our approach emphasizes reducing energy costs and carbon emissions while ensuring privacy. Experimental results demonstrate that Federated TRPO is comparable with state-of-the-art federated RL methodologies without hyperparameter tuning. The proposed framework highlights the feasibility of collaborative learning for achieving optimal control policies in energy systems, advancing the goals of sustainable and efficient smart grids. Our code is accessible in *this repo*.

1. Introduction

Global reviews of progress under the SDGs and the 2015 Paris Agreement still show the world drifting beyond the 1.5–2 °C ceiling, as confirmed by the latest IPCC assessment (IPCC, 2023). Because building operations consume about 30% of global final energy and produce roughly 26% of energy-related CO_2 emissions (International Energy Agency, 2024), cutting their footprint is pivotal. A recent analysis estimates that AI-enabled controls alone could trim building energy demand and emissions by 8–19% by 2050 (Ding et al., 2024). Machine learning (ML) is a technology that, when used correctly, can accelerate the actions needed to fight global warming. ML can support the transition from fossil fuels, accelerate renewable energy adoption, improve energy efficiency, and promote sustainable transportation, offering powerful tools to meet some of the most pressing climate goals (Rolnick et al., 2019).

Rooftop solar and other on-site renewables introduce fast power swings that can create acute shortage events at the building level (Zheng et al., 2024; Perera et al., 2020; Gernaat et al., 2021). Modern Building Energy Management Systems (BEMS) mitigate these fluctuations by pairing local storage with predictive control and have been shown to cut energy use and costs across diverse facilities (Michailidis et al., 2024). Model-free reinforcement learning (RL) can learn such control policies directly from data, capturing the non-linear dynamics of distributed resources without a detailed physics model (Barbalho et al., 2025). Yet standard RL is sample-hungry and generalises poorly to unseen climates or occupancy patterns (Bekal et al., 2025). Privacy-preserving federated approaches overcome both issues by exchanging model updates rather than raw traces, protecting occupant data while pooling experience across sites (Lee et al., 2021). Recent demonstrations report cost and CO_2 reductions when scaling federated RL to dozens of heterogeneous buildings in multiple climates (Sievers et al., 2025a). Privacy guarantees remove regulatory and organisational barriers to pooling knowledge across many buildings, making collaborative learning a practical path to scalable net-zero control (Sievers et al., 2025b).

This study explores building-scale net-zero microgrids ("a group of interconnected loads and distributed energy resources within clearly defined electrical boundaries that acts as a single controllable entity" (Ton & Smith, 2012)) as test-beds for scalable, privacy-preserving control. We generated scenarios where buildings are self-sustainable by design as long as the electrical storage is managed using the optimal policy. Using a custom microgrid scenario based on CityLearn (Vázquez-Canteli et al., 2019), we explored the generalization capabilities of RL agents and the challenges they faced even in simplified scenarios showing that naive policies are difficult to learn in a collaborative way. We improved the learning process using Federated Learning (FL), in this setup each building trains a local RL agent but only shares parameter updates, a practice that retains

[*]Equal contribution [1]Department of Machine Learning, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Correspondence to: Nicolas Cuadrado <nicolas.avila@mbzuai.ac.ae>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

data privacy while exploiting the scalability and generalisation benefits identified in recent federated-learning surveys (Kairouz et al., 2021; Lee et al., 2021). Additionally, we included personalisation through split-learning (Han et al., 2022), further shrinking the generalisation gap that often appears when controllers are deployed in heterogeneous conditions or occupancy patterns (Bekal et al., 2025). We present a summary of our contributions next:

- We introduce a custom environment based on CityLearn, which allows a better understanding of the learned policies and the impact of the reward function in the learning process. We also modified some characteristics that better represent the real-life dynamics.
- We provide experimental assessment of traditional RL algorithms in our environment, testing generalization and optimality of the learned policies.
- Our results demonstrate the generalization capabilities of FL and how reasonable considerations like personalization and split-learning (Han et al., 2022) can reduce the optimality gap in the learned policies.
- We provide a wrapper for the CityLearn environment that makes it compatible with TorchRL (Bou et al., 2023), improving the modularity and portability of our code. We also provided a version of the TRPO algorithm using TorchRL.

2. Related Work

Federated Learning. FL trains a common model by letting each client update locally and then merging weights. The seminal FedAvg algorithm performs this aggregation on a central server (McMahan et al., 2017), whereas Gossip protocols average peer-to-peer without a server (Koloskova et al., 2019). Subsequent work addresses FedAvg's limits under non-IID data and privacy leakage: split learning keeps part of each network private to enable personalisation (Vepakomma et al., 2018), and differential privacy injects calibrated noise into updates (Geyer et al., 2017). Studies on heterogeneous clients (Diao et al., 2021) and generalisation theory (Yuan et al., 2022) confirm that, with these extensions, FL can learn robust, privacy-aware controllers—making it well suited to smart-grid scenarios with diverse buildings.

RL-based and Federated RL-based microgrids. Recent studies model microgrids as groups of houses with distinct loads, rooftop renewables, and storage devices, typically using the open-source CityLearn platform (Vázquez-Canteli et al., 2019; Nweye et al., 2024) built on Gymnasium (Towers et al.) and real demand data (Wilson et al., 2021). Within CityLearn, (Nweye et al., 2023) trained multi-agent Soft Actor-Critic controllers for battery systems to cut cost, carbon, and peak demand, while (Nweye et al., 2022a) showed that warm-starting RL with rule-based policies speeds convergence. Beyond public benchmarks, a model-based RL approach with constrained optimisation targets multi-energy hubs (Ceusters et al., 2021); a federated RL scheme addresses free-riders and non-IID data (Su et al., 2022); and hierarchical FRL embeds FedAvg in layered control architectures (Cuadrado et al., 2023a;b). Collectively, these works highlight both the promise and the outstanding challenges of RL and FRL for collaborative, net-zero microgrids.

Despite steady progress, two gaps remain. First, many studies treat a monotonic rise in cumulative reward as proof of optimality, yet trivial local minima, such as always charging or doing nothing, can produce similar curves. Second, CityLearn challenge winners (2021–2023) optimised the highest-weighted terms and largely ignored CO_2 impact (Nweye et al., 2022b; Nagy et al., 2021), even though emission reduction should be central in climate-aligned control.

3. Method

3.1. Environment

We used our customized version of the CityLearn environment. We decided to generate synthetic data that allows us to understand clearly the learned policies through simple cases, focusing on representing homes that can be self-sustainable, in other words, each home can provide enough energy to serve its demand, so the optimal solution must have zero cost and emissions. Our data generation process takes as a reference the *citylearn_challenge_2022_phase_all* scenario. It modifies the time series for *non-shiftable load* to ensure solar energy can cover the demand profile completely. We included other modifications to the environment:

- New logic for the selling energy rate. We found it unrealistic that in the original scenario, the users receive compensation equal to the Time-of-Use (ToU) tariff. This setup discourages houses from being self-sustainable.
- We simplified the storage models so they don't suffer degradation or decay in their performance, helping to ease the understanding of the policies we are learning, which is one of the main interests of this research.
- We introduced additional observations representing load predictions for some steps ahead.

We kept the grid price pattern in the original CityLearn environment, which fluctuates throughout the day, peaking during high-demand hours, while the solar panels contribute energy during daylight hours. Allowing the agent to stockpile energy during off-peak times and, when solar generation exceeds immediate needs, later release stored energy during high-price periods. Emissions follow the actual emission rate from the grid mix (kgCO2e/kWh) following the original dataset used in CityLearn. We generate convex combinations of the training buildings' time series for the validation

data. Each episode spans twenty-four hours, and t is the index for the time step. For a building i, an agent considers the accessible observations to decide the charge or discharge of the battery in a time step ($E_{t,i}^{\text{batt}}$), aiming to meet the building's load requirements. Its primary objective is minimizing its house's overall energy cost and emissions and, as an extension, for the whole microgrid. The table 3 in the Appendix presents an overview of the RL case. Our design of the reward function is a weighted sum of the resulting cost and emissions for the complete episode (a full day), and we established the values as $W_C = 0.4$ and $W_G = 0.6$ to enforce the prioritization of emission reduction. We also added a penalization to the reward function to guide the agent to avoid unfeasible actions.

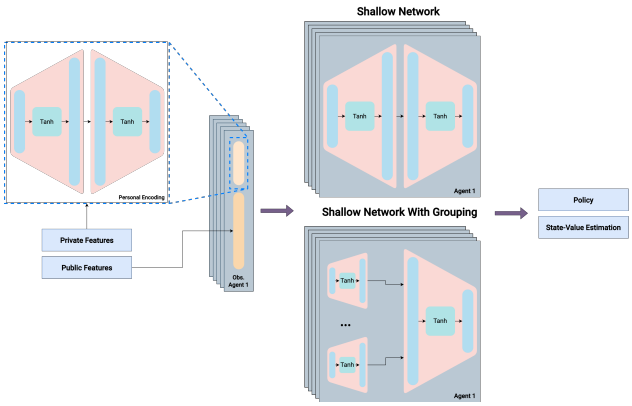


Figure 1. Overview of our models: We use the same type of network for Policy and State-Value networks, changing only the output layer. The *Personal Encoding* block is optional.

3.2. Model

We explore the advantages of a collaborative learning approach for the case of microgrids, given their distributed nature: each household has an agent that explores and learns from different sections of the state space, given the individual attributes of the household (generation, load profile, battery capacity). We have two types of models for both Policy and Critic networks: a shallow neural network and a custom model that adds a grouping feature logic as the first part of the feature extraction; we refer to it as *Personal Encoding* in Figure 1. This block represents a part of the network that remains local; the FedAvg process does not access it. It encodes features that should remain private (e.g., number of people living in the household, ToU strategy, appliances, area of the house, etc.) and learns a latent representation, used as an input to the public model, stacked with the public features, which aligns with the concept of split-learning (Vepakomma et al., 2018). We can assume it as a personalization strategy.

4. Experiments

4.1. Experiment Setup

We generated two environments for our experiments using the simplified version of the data 3(a): a microgrid with two houses and one with five houses. Each one also has a shifted version that creates day gaps between the house's time series, simulating that each house is subject to different environmental conditions and solar irradiance patterns. Regarding the model configurations, as described 1, we run experiments in four setups: shallow network, shallow network with personal encoding, shallow network with grouping, and shallow network with personal encoding and grouping. We defined four group features: load-related, solar-related, pricing-related, and the rest. We fixed the communication rounds for both algorithms we tried, Proximal Policy Optimisation(PPO) and Trust Region Policy Optimisation (TRPO), to evaluate how close they can come to the optimal solution within the same number of iterations. Another essential detail is setting the discount factor $\gamma = 1$, meaning that our agents don't consider a discounted reward, given that we are running episodes of just one day (24 time steps).

4.2. Quantitative Results

The tables 1 and 2 show the experiments we mentioned in the previous section: *base* means we used the shallow network, *pe* that we included personal encoding, and *gf* that we grouped features. We ran each combination for five different seeds. We averaged the results, making 160 experiments with an average run time of $70m$ for the PPO runs, and $110m$ for the TRPO runs using an NVIDIA RTX 4000 Ada Generation GPU.

Starting with the environment containing just two buildings, we found that training with the shifted variant is more challenging than the normal one, which makes sense considering we provide more diverse data without increasing the minibatch size. We cannot conclude from the results that using personal encoding or grouped features made a difference in the case of PPO, but for TRPO, it seems that they could be beneficial, as was evident in the shifted dataset. In general, as expected, the performance of the tuned PPO was superior to that of the TRPO, and the reward function value demonstrated it.

In the five-building environment, grouping features seems to play a more critical role than in the case of just having two buildings. We interpret this as the effect of having more agents, which means more sampling and more diverse data, which can help the networks unveil patterns in the data that are not evident by simple inspection. However, this does not seem to be the case for TRPO, in which the simple model yields the best performance, although it is essential to note that the performance of TRPO is poor based on the reward

Table 1. **Experimental Results for 2-Building Environment:** Values in the table represent the average for 5 seeds. Values in parentheses represent the difference concerning not having battery management.

	Experiment	Train Cost	Eval Cost	Train Emissions	Eval Emissions	Train Reward	Eval Reward
PPO	base	0.04227 (-0.08977)	0.04099 (-0.09148)	0.06380 (-0.07487)	0.06260 (-0.07652)	-0.19331	-0.18754
	gf	0.03054 (-0.10150)	**0.03049 (-0.10199)**	0.04341 (-0.09526)	**0.04308 (-0.09604)**	-0.13227	**-0.13188**
	pe	0.03885 (-0.09319)	0.04630 (-0.08618)	0.05841 (-0.08026)	0.06384 (-0.07527)	-0.17686	-0.19956
	pe gf	**0.03051 (-0.10153)**	0.03227 (-0.10021)	**0.04306 (-0.09561)**	0.04637 (-0.09274)	**-0.13065**	-0.14021
	base shifted	0.04923 (-0.09717)	**0.07062 (-0.07229)**	0.08726 (-0.10790)	**0.11533 (-0.08125)**	-0.22554	**-0.34312**
	gf shifted	0.24691 (+0.10051)	0.25710 (+0.11419)	0.28074 (+0.08558)	0.30601 (+0.10943)	-1.46458	-1.51916
	pe shifted	**0.04541 (-0.10099)**	0.10654 (-0.03637)	**0.08211 (-0.11305)**	0.13600 (-0.06058)	**-0.21291**	-0.56921
	pe gf shifted	0.10324 (-0.04316)	0.15339 (+0.01049)	0.10922 (-0.08594)	0.17679 (-0.01979)	-1.10813	-1.34919
TRPO	base	0.31658 (+0.18454)	0.32391 (-0.19143)	0.26502 (+0.12635)	0.26908 (+0.12997)	-1.12509	-1.12524
	gf	0.31882 (+0.18678)	0.32288 (+0.19040)	0.27001 (+0.13134)	0.27020 (+0.13108)	-1.02230	-1.01532
	pe	0.10567 (-0.02637)	0.10402 (-0.02846)	**0.11511 (-0.02357)**	**0.11271 (-0.02641)**	**-0.47736**	**-0.47064**
	pe gf	**0.10519 (-0.02685)**	**0.10355 (-0.02893)**	0.11519 (-0.02348)	0.11302 (-0.02609)	-0.47877	-0.47164
	base shifted	0.15482 (+0.00842)	0.14608 (+0.00317)	0.18679 (-0.00837)	0.17856 (-0.01802)	-0.61017	-0.58478
	gf shifted	0.17848 (0.03208)	0.16951 (+0.02660)	0.19592 (+0.00076)	0.18621 (-0.01037)	-0.77824	-0.74692
	pe shifted	0.14833 (+0.00193)	0.14171 (-0.00120)	0.17103 (-0.02413)	0.16301 (-0.03357)	-0.64225	-0.63830
	pe gf shifted	**0.12728 (-0.01912)**	**0.12032 (-0.02259)**	**0.15287 (-0.04229)**	**0.14513 (-0.05145)**	**-0.57360**	**-0.56441**

Table 2. **Experimental Results for 5-Building Environment:** Values in the table represent the average for 5. Values in parentheses represent the difference concerning not having battery management.

	Experiment	Train Cost	Eval Cost	Train Emissions	Eval Emissions	Train Reward	Eval Reward
PPO	base	0.04001 (-0.09140)	0.03856 (-0.09297)	0.06094 (-0.07710)	0.05949 (-0.07866)	-0.18392	-0.17930
	gf	0.24239 (+0.11098)	0.25020 (+0.11867)	0.23766 (+0.09962)	0.24276 (+0.10461)	-1.55780	-1.55296
	pe	0.04089 (-0.09052)	0.04505 (-0.08648)	0.06159 (-0.07645)	0.06488 (-0.07328)	-0.18951	-0.20773
	pe gf	**0.02889 (-0.10253)**	**0.03233 (-0.09920)**	**0.04235 (-0.09569)**	**0.04711 (-0.09104)**	**-0.13100**	**-0.14662**
	base shifted	0.05985 (-0.09039)	0.06716 (-0.08331)	0.08469 (-0.07876)	0.09063 (-0.07437)	-0.26297	-0.28186
	gf shifted	**0.03725 (-0.11299)**	**0.05633 (-0.09414)**	**0.05421 (-0.10924)**	**0.07025 (-0.09474)**	**-0.16773**	**-0.22911**
	pe shifted	0.06305 (-0.08719)	0.08847 (-0.06200)	0.09080 (-0.07265)	0.10826 (-0.05674)	-0.28214	-0.39795
	pe gf shifted	0.10324 (-0.04316)	0.15339 (+0.01049)	0.10922 (-0.08594)	0.17679 (-0.01979)	-1.10813	-1.34919
TRPO	**base**	**0.09811 (-0.03330)**	**0.09737 (-0.03416)**	0.11173 (-0.02630)	0.11086 (-0.02730)	**-0.46143**	-0.45768
	gf	0.10437 (-0.02704)	0.10312 (-0.02841)	0.11455 (-0.02349)	0.11327 (-0.02488)	-0.47631	-0.47153
	pe	0.09843 (-0.03298)	0.09750 (-0.03403)	**0.11019 (-0.02785)**	**0.10906 (-0.02909)**	-0.46143	**-0.45614**
	pe gf	0.10177 (-0.02964)	0.10044 (-0.03109)	0.11274 (-0.02529)	0.11140 (-0.02676)	-0.46460	-0.45958
	base shifted	0.13025 (-0.01999)	**0.11929 (-0.03118)**	**0.12936 (-0.03409)**	**0.12229 (-0.04270)**	-0.57766	**-0.53850**
	gf shifted	0.17795 (+0.02771)	0.16457 (+0.01410)	0.16254 (-0.00091)	0.15295 (-0.01204)	-0.77981	-0.74353
	pe shifted	**0.12846 (-0.02178)**	0.11979 (-0.03068)	0.13076 (-0.03269)	0.12391 (-0.04109)	**-0.56240**	-0.55708
	pe gf shifted	0.14833 (+0.00193)	0.14171 (-0.00120)	0.17103 (-0.02413)	0.16301 (-0.03357)	-0.64225	-0.63830

function. PPO overperforms TRPO significantly, but we highlight that we didn't spend time fine-tuning it. Still, it managed to identify some patterns in the data, as we will show in the qualitative results in the Appendix.

5. Conclusion

We examined battery control in net-zero microgrids. Using a stripped-down CityLearn scenario with a known optimal policy, we studied RL combined with FedAvg and split-learning personalisation to balance privacy, heterogeneity, and policy quality. Federated TRPO performed on par with a tuned federated PPO: PPO was sometimes superior but stalled near the optimum, whereas TRPO converged reliably without hyperparameter tuning. A pragmatic remedy is to warm-start with a few PPO iterations and then switch to TRPO, leveraging TRPO's adaptive step size in the final optimisation phase. The study shows that even "simple" optimal policies are surprisingly hard to learn, supporting curriculum designs that start with well-controlled cases before moving to full real-world complexity.

Future Work. Next steps include benchmarking against model-predictive control (MPC) and additional RL baselines to gauge relative sample-efficiency and robustness. We will design new "toy" microgrid cases—varying storage sizes, demand shapes, and renewable mixes—to probe policy behaviour and measure the optimality gap more precisely. Finally, we plan to scale experiments from five to dozens of buildings, testing how privacy, communication cost, and convergence behave in larger federations.

References

Barbalho, P. I. N., Moraes, A. L., Freitas, V. A. L., Barra, P. H. A., Fernandes, R. A. S., and Coury, D. V. Reinforcement learning solutions for microgrid control and management: A survey. *IEEE Access*, 13:1–35, 2025. doi: 10.1109/ACCESS.2025.3546578.

Bekal, G. U., Ghareeb, A., and Pujari, A. Continual reinforcement learning for hvac systems control: Integrating hypernetworks and transfer learning. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2503.19212.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Bou, A., Bettini, M., Dittert, S., Kumar, V., Sodhani, S., Yang, X., Fabritiis, G. D., and Moens, V. Torchrl: A data-driven decision-making library for pytorch, 2023. URL https://arxiv.org/abs/2306.00577.

Ceusters, G., Rodríguez, R. C., García, A. B., Franke, R., Deconinck, G., Helsen, L., Nowé, A., Messagie, M., and Camargo, L. R. Model-predictive control and reinforcement learning in multi-energy system case studies. *CoRR*, abs/2104.09785, 2021. URL https://arxiv.org/abs/2104.09785.

Cuadrado, N. M., Guillén, R. A. G., and Takác, M. FRESCO: federated reinforcement energy system for cooperative optimization. In Maughan, K., Liu, R., and Burns, T. F. (eds.), *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/pdf?id=75mWq5j4iso.

Cuadrado, N. M., Gutiérrez, R. A., Zhu, Y., and Takác, M. MAHTM: A multi-agent framework for hierarchical transactive microgrids. *CoRR*, abs/2303.08447, 2023b. doi: 10.48550/arXiv.2303.08447. URL https://doi.org/10.48550/arXiv.2303.08447.

Diao, E., Ding, J., and Tarokh, V. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=TNkPBBYFkXg.

Ding, C., Ke, J., Levine, M., Zhou, N., et al. Potential of artificial intelligence in reducing energy and carbon emissions of commercial buildings at scale. *Nature Communications*, 15(5916), 2024. doi: 10.1038/s41467-024-50088-4.

Gernaat, D. E. H. J., de Boer, H. S., Daioglou, V., Yalew, S. G., Müller, C., and van Vuuren, D. P. Climate change impacts on renewable energy supply. *Nature Climate Change*, 11(2):119–125, Feb 2021. ISSN 1758-6798. doi: 10.1038/s41558-020-00949-9. URL https://doi.org/10.1038/s41558-020-00949-9.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *CoRR*, abs/1712.07557, 2017. URL http://arxiv.org/abs/1712.07557.

Han, D., Kim, D., Choi, M., Brinton, C. G., and Moon, J. Splitgp: Achieving both generalization and personalization in federated learning. *CoRR*, abs/2212.08343, 2022. doi: 10.48550/arXiv.2212.08343. URL https://doi.org/10.48550/arXiv.2212.08343.

International Energy Agency. Buildings – energy system, 2024. URL https://www.iea.org/energy-system/buildings. Accessed: 7 June 2025.

IPCC. Climate change synthesis report. https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_SPM.pdf, 2023. (Accessed on 09/13/2024).

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., and et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210, 2021. doi: 10.1561/2200000083.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Koloskova, A., Stich, S. U., and Jaggi, M. Decentralized stochastic optimization and gossip algorithms with compressed communication. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3478–3487. PMLR, 2019. URL http://proceedings.mlr.press/v97/koloskova19a.html.

Lee, S., Xie, L., and Choi, D.-H. Privacy-preserving energy management of a shared energy storage system for smart buildings: A federated deep reinforcement learning approach. *Sensors*, 21(14):4898, 2021. doi: 10.3390/s21144898.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In

Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

Michailidis, P., Michailidis, I., and Kosmatopoulos, E. Review and evaluation of multi-agent control applications for energy management in buildings. *Energies (19961073)*, 17(19), 2024.

Nagy, Z., Vázquez-Canteli, J. R., Dey, S., and Henze, G. The citylearn challenge 2021. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '21, pp. 218–219, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391146. doi: 10.1145/3486611.3492226. URL https://doi.org/10.1145/3486611.3492226.

Narayanamurthy, R., Handa, R., Tumilowicz, N., Herro, C., and Shah, S. Grid integration of zero net energy communities. *ACEEE Summer Study Energy Effic. Build*, 2016.

Nweye, K., Liu, B., Stone, P., and Nagy, Z. Real-world challenges for multi-agent reinforcement learning in grid-interactive buildings. *Energy and AI*, 10:100202, 2022a.

Nweye, K., Nagy, Z., Mohanty, S., Chakraborty, D., Sankaranarayanan, S., Hong, T., Dey, S., Henze, G., Drgona, J., Lin, F., et al. The citylearn challenge 2022: Overview, results, and lessons learned. *NeurIPS 2022 Competition Track*, pp. 85–103, 2022b.

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

Nweye, K., Kaspar, K., Buscemi, G., Fonseca, T., Pinto, G., Ghose, D., Duddukuru, S., Pratapa, P., Li, H., Mohammadi, J., Lino Ferreira, L., Hong, T., Ouf, M., Capozzoli, A., and Nagy, Z. Citylearn v2: energy-flexible, resilient, occupant-centric, and carbon-aware management of grid-interactive communities. *Journal of Building Performance Simulation*, pp. 1–22, November 2024. ISSN 1940-1507. doi: 10.1080/19401493.2024.2418813. URL http://dx.doi.org/10.1080/19401493.2024.2418813.

Perera, A. T. D., Nik, V. M., Chen, D., Scartezzini, J.-L., and Hong, T. Quantifying the impacts of climate change and extreme climate events on energy systems. *Nature Energy*, 5(2):150–159, Feb 2020. ISSN 2058-7546. doi: 10.1038/s41560-020-0558-0. URL https://doi.org/10.1038/s41560-020-0558-0.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Körding, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J. T., and Bengio, Y. Tackling climate change with machine learning. *CoRR*, abs/1906.05433, 2019. URL http://arxiv.org/abs/1906.05433.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Sievers, J., Henrich, P., Beichter, M., Mikut, R., Hagenmeyer, V., Blank, T., and Simon, F. Federated reinforcement learning for sustainable and cost-efficient energy management. *Energy and AI*, 21:100521, 2025a. doi: 10.1016/j.egyai.2025.100521.

Sievers, J., Henrich, P., Beichter, M., Mikut, R., Hagenmeyer, V., Blank, T., and Simon, F. Federated reinforcement learning for sustainable and cost-efficient energy management. *Energy and AI*, 21:100521, 2025b. doi: 10.1016/j.egyai.2025.100521.

Su, Z., Wang, Y., Luan, T. H., Zhang, N., Li, F., Chen, T., and Cao, H. Secure and efficient federated learning for smart grid with edge-cloud collaboration. *IEEE Trans. Ind. Informatics*, 18(2):1333–1344, 2022. doi: 10.1109/TII.2021.3095506. URL https://doi.org/10.1109/TII.2021.3095506.

Ton, D. T. and Smith, M. A. The u.s. department of energy's microgrid initiative. *The Electricity Journal*, 25(8):84–94, 2012. ISSN 1040-6190. doi: https://doi.org/10.1016/j.tej.2012.09.013. URL https://www.sciencedirect.com/science/article/pii/S1040619012002254.

Towers, M., Kwiatkowski, A., Terry, J. K., Balis, J. U., de Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Tan, H. J. S., and Younis, O. G. Gymnasium: A Standard Interface for Reinforcement Learning Environments. URL https://github.com/Farama-Foundation/Gymnasium.

Vázquez-Canteli, J. R., Kämpf, J. H., Henze, G., and Nagy, Z. Citylearn v1.0: An openai gym environment for demand response with deep reinforcement learning. In *Proceedings of the 6th ACM International*

Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys 2019, New York, NY, USA, November 13-14, 2019, pp. 356–357. ACM, 2019. doi: 10.1145/3360322.3360998. URL https://doi.org/10.1145/3360322.3360998.

Vepakomma, P., Gupta, O., Swedish, T., and Raskar, R. Split learning for health: Distributed deep learning without sharing raw patient data. *CoRR*, abs/1812.00564, 2018. URL http://arxiv.org/abs/1812.00564.

Wilson, Eric, Parker, Andrew, Fontanini, Anthony, Present, Elaina, Reyna, Janet, Adhikari, Rajendra, Bianchi, Carlo, CaraDonna, Christopher, Dahlhausen, Matthew, Kim, Janghyun, LeBar, Amy, Liu, Lixi, Praprost, Marlena, White, Philip, Zhang, Liang, DeWitt, Peter, Merket, Noel, Speake, Andrew, Hong, Tianzhen, Li, Han, Frick, M., Natalie, Wang, Zhe, Blair, Aileen, Horsey, Henry, Roberts, David, Trenbath, Kim, Adekanye, Oluwatobi, Bonnema, Eric, Kontar, E., Rawad, Gonzalez, Jonathan, Horowitz, Scott, Jones, Dalton, Muehleisen, Ralph, Platthotam, Siby, Reynolds, Matthew, Robertson, Joseph, Sayers, Kevin, , and Li, Q. End-use load profiles for the u.s. building stock, 10 2021. URL https://data.openei.org/submissions/4520.

Yuan, H., Morningstar, W. R., Ning, L., and Singhal, K. What do we mean by generalization in federated learning? In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=VimqQq-i_Q.

Zheng, D., Tong, D., Davis, S. J., Qin, Y., et al. Climate change impacts on the extreme power shortage events of wind–solar supply systems worldwide during 1980–2022. *Nature Communications*, 15(5225), 2024. doi: 10.1038/s41467-024-48966-y.

A. Definition of RL Scenario

A.1. Reinforcement Learning

The RL component of our methods happens at the local training. We compared TRPO against a state-of-the-art algorithm called Proximal Policy Optimization (PPO) (Schulman et al., 2017), which emerged as an alternative to TRPO that avoids the use of second-order information. PPO relies on the same theoretical base as TRPO but redefines the policy optimization objective as an unconstrained expression. The scenario we study in this work, and the case of microgrids' agents, does not require the use of very deep neural networks, and that is a reason why we believe that second-order methods are feasible in these cases, especially for RL methods, which tend to be very sensitive to the learning rate. TRPO explores the vast space by identifying suitable step sizes within a confined region, ensuring stable policy, monotonic improvement, and more effective exploration than other policy gradient methods.

Another essential aspect of opting for an RL approach is defining a proper reward function, especially considering that we defined a continuous action space, which means the number of possible actions is unbounded. The typical approach for this case is to use a policy network that outputs the parameters of a Gaussian distribution (reparameterization trick) to sample a real value that defines the action from that distribution. Throughout the training process, we don't have guarantees that the policy network will explore feasible actions (within the range specified in 3), especially in the first iterations. That fact motivated the addition of a penalization $P_{i,t}$ that impacts the return of a trajectory (the accumulated sum or rewards.) An essential consideration we made when designing our reward function is that we are not guiding it to follow a specific policy we know beforehand; we are just guiding the exploration of the agent by avoiding ineffective actions.

A.2. Federated RL

An actual FL setup implies having clients interact through a communication network channel throughout the training rounds. In the context of our study case, we can imagine it as the agents operating in an embedded device that controls the batteries of each household and connects to a parameter server through a dedicated network connection. However, this case would introduce some of the challenges of FL: limited bandwidth, faulty nodes or updates, Byzantine clients, etc. The main interest of this work is not to alleviate those challenges but to explore the benefits of collaborative learning. Thus, we are not required to implement the exact flow of FL to perform our experiments.

As Figure 1 shows, we stack the features (observations and personal encoding) for all the agents and pass them to their corresponding networks. Passing the stacked observations to a single network has the same effect as doing a FedAvg: each agent samples their trajectories, and after a forward pass, the backward pass will propagate the average gradient of all of them, which is, in practice, the same behavior we expect from the FL methodology. We use a simple one-hot encoder for each home as the personal encoding. Although we are not in the process of training a private part of the policy network, we can explore the exact effects of personalization by having a vector representation that identifies each home. Using FL is to discern the interconnected relationships in the observations across buildings collectively. Relying solely on data from a single building risks overfitting, but leveraging diverse data distributions from multiple buildings helps uncover latent connections. Moreover, the model adeptly extrapolates unseen data instances by assimilating insights from disparate building data distributions, handling anomalies emerging from using usage patterns.

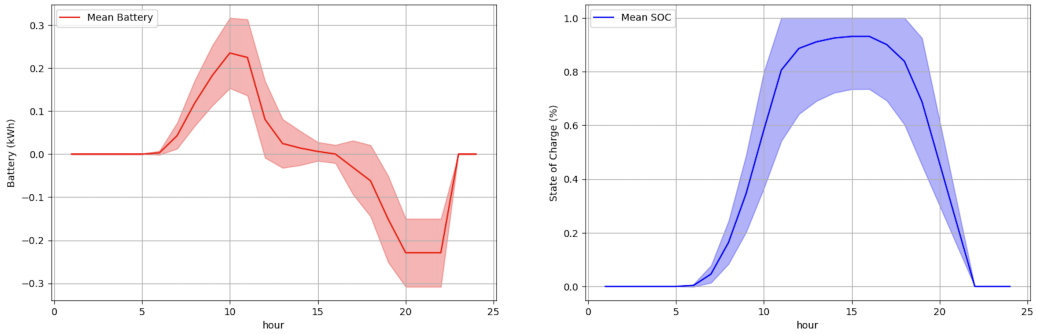
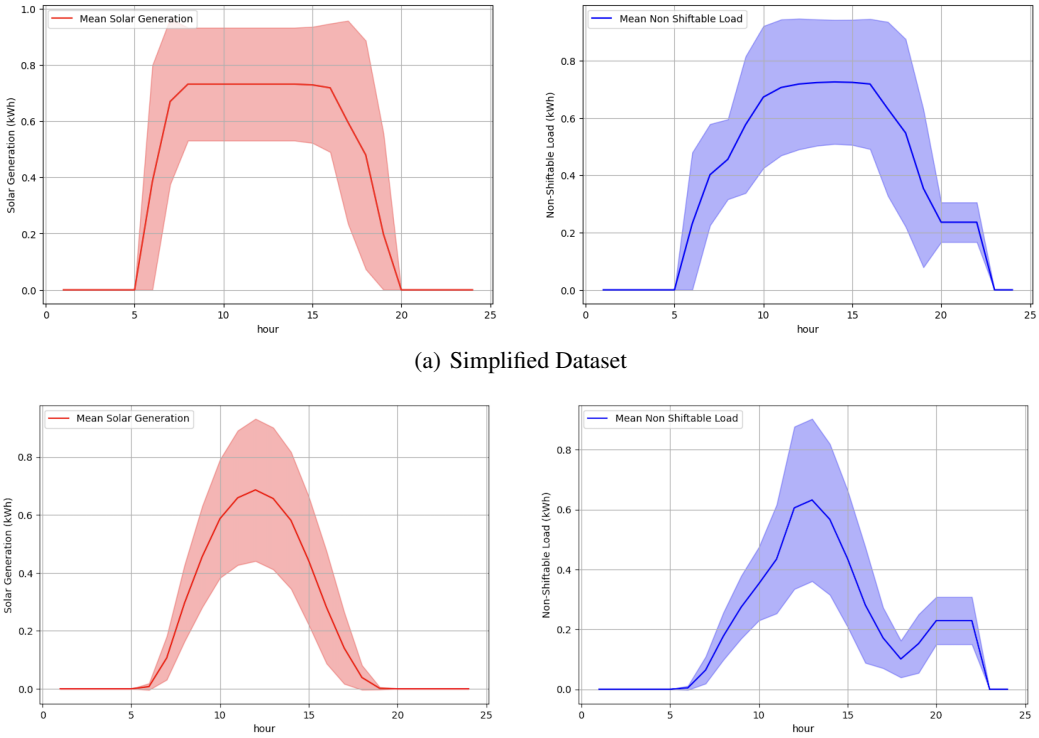


Figure 2. Optimal Actions and State of Charge: Mean and variance throughout the day, considering a complete year, for a building. By design, we know the optimal policy that must yield zero emissions and cost beforehand.

B. Synthetic Data

For this research, we decided to generate our data as it allowed us to explain better the policies we learned. Although using accurate data is always desirable for ML, there is low availability, especially regarding energy usage patterns, which concerns privacy. For example, the CityLearn environment constructs its 2022 scenario using data from 17 Sierra Crest Zero Net Energy community buildings from another study about ToU strategies (Narayanamurthy et al., 2016). They couldn't access all the accurate data details, so they had to generate data that emulated the behavior based on some prior information about the ToU strategies the households might have used. Research about microgrid control systems is not particularly affected by lack of data, as prior information about consumption behaviors, typical weather conditions given a location, seasonal changes, and factors influencing energy use are accessible. Synthetic data could represent typical scenarios an agent may face and have the advantage of creating relevant evaluation scenarios by adding noise or making a convex combination of seen data.

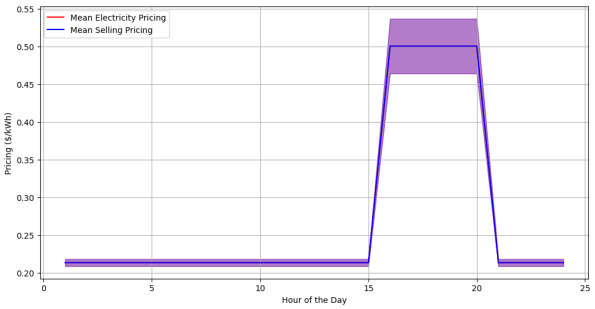


(a) Simplified Dataset

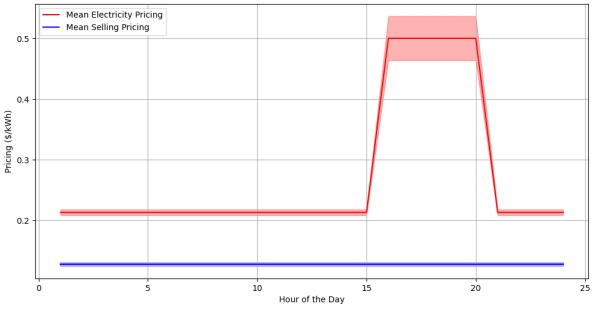
(b) Normal Dataset

Figure 3. Sample of the Mean and Variance of Solar Generation and Non-Shiftable Load over 24 hours for a year of data for one of the buildings.

We generated our synthetic data by altering the Non-Shiftable Load time series using a simple rule that defines the policy we want to learn: the households have enough energy and battery capacity to serve their demand when there is solar energy and to store what is necessary when there is no sun. By definition, our scenario leads to net-zero energy consumption. Figure 3 shows how two versions of the data we generated, in the case of 3(a), we simplified the solar profile by making it constant through the sunny hours. For 3(b), we kept the solar profile as in the original reference case from CityLearn. By simplifying these profiles, we intended to reduce the state space for the policy while still keeping it challenging to learn. Another critical difference in our scenario is the pricing for energy. Our reference scenario from the CityLearn environment used the real pricing time series used in the Zero Crest Community Study. Still, for the environment implementation, they defined that the selling price is the same tariff whenever a household sends back energy to the grid. By inspecting 4(a), it is easy to notice that a policy optimizing exclusively by the energy cost would charge as much as possible in the low-price hours to sell energy when it is almost twice the buying price. In our scenario, we defined a margin of 40% of the lowest daily electricity pricing to ensure we don't learn policies that bypass the weighting factor W_G by finding profitable energy trades.



(a) Original scenario



(b) Our scenario

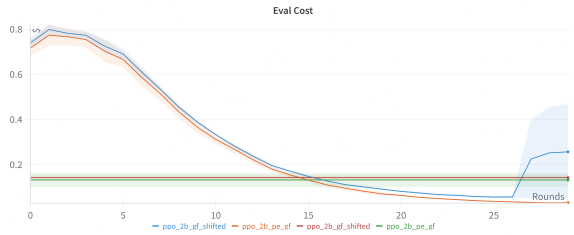
Figure 4. Mean and Variance of Electricity Pricing (ToU) and the Selling Pricing over 24 hours for a year of data.

C. Hyperparameters

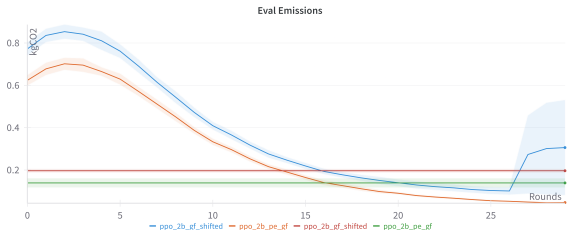
We perform a hyperparameter search with Wandb sweeps (Biewald, 2020), using Bayesian search to find the best configurations for PPO and TRPO with the following final configurations:

C.0.1. Quantitative Results

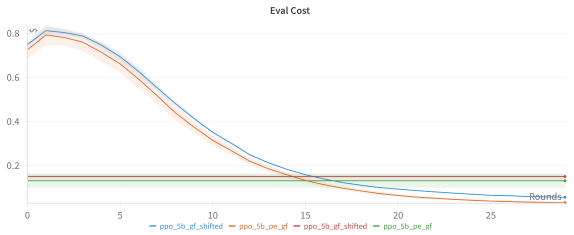
We inspected the best-performing policies, considering the eval. reward, according to Tables 1 and 2 to understand if we accomplished the objective of beating the baseline policy, which is not using a battery. In the case of PPO, we can notice in 5 that it was the case. However, we can see an effect we mentioned before in Figures 5(a) and 5(b): extending the training to reduce the optimality gap destabilizes the training, which is something that happens to gradient-based optimizers with a fixed learning rate.



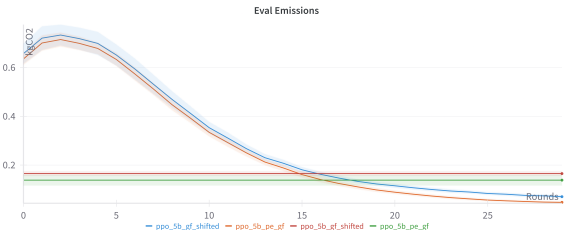
(a) 2B Eval. Energy Cost



(b) 2B Eval. Emissions



(c) 5B Eval. Energy Cost

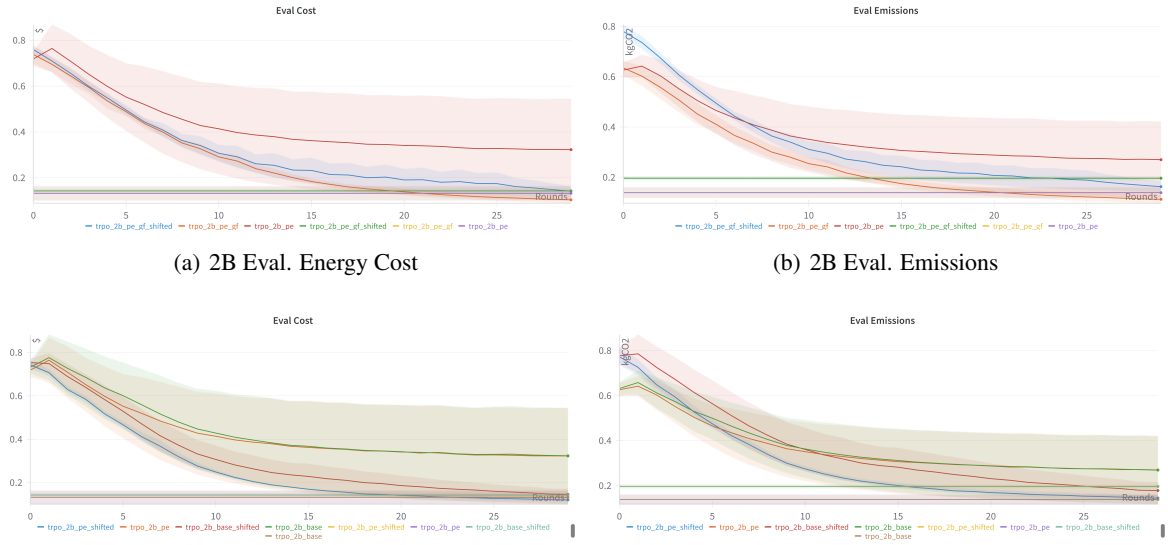


(d) 5B Eval. Emissions

Figure 5. **Best PPO policies:** Plots show mean with an envelope of one standard deviation computed from the five different seeds.

Increasing the number of buildings makes it harder for the agents to approach net-zero energy and training with the shifted version of the dataset, something expected. A first inspection of the results for the five-building environment might give the impression that the training process was more stable, but the reason must be that the policy didn't get as close to net zero as in the case of two buildings; the instability in the training happens in the vicinity of the optimal solution.

For the best-performing policies using TRPO, we can notice how the low returns reflect high variance among runs. However,

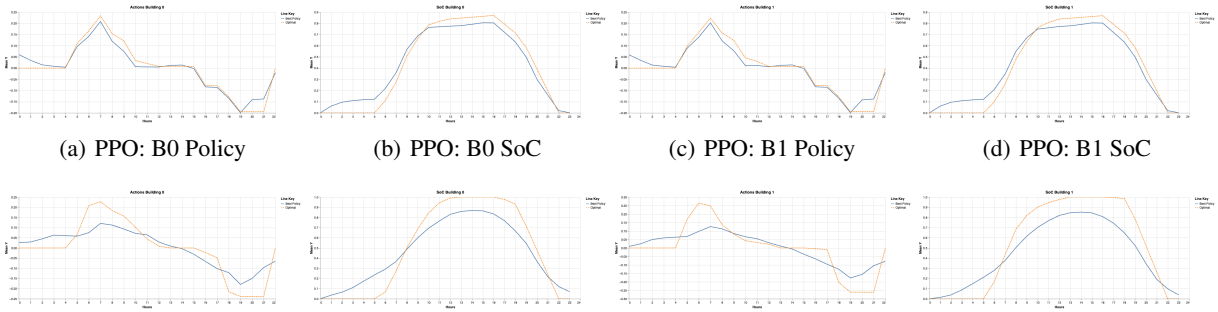


(a) 2B Eval. Energy Cost (b) 2B Eval. Emissions

(c) 5B Eval. Energy Cost (d) 5B Eval. Emissions

Figure 6. **Best TRPO policies:** Plots show mean with an envelope of one standard deviation computed from the five different seeds.

some configurations (model with personal encoding and grouping) limited the variance. They beat the baseline on each run, which is a highlight considering the lack of tuning. That event was present in both environments.



(a) PPO: B0 Policy (b) PPO: B0 SoC (c) PPO: B1 Policy (d) PPO: B1 SoC

(e) TRPO: B0 Policy (f) TRPO: B0 SoC (g) TRPO: B1 Policy (h) TRPO: B1 SoC

Figure 7. **Best policies for the 2-building environment:** Comparison of learned policy against the optimal policy, averaged for the five seeds.

In Figure 7, we notice that PPO gets very close to the optimal policy, but a few misalignments explain why it couldn't reach net zero. From the best TRPO policy, we can highlight that although it's far from optimal, it is evident that it learned a pattern, and that although sub-optimal, it's still a policy that beats the baseline.

D. Results Summary

When analyzing the results, it is essential to consider that the PPO approach was tuned and used the best set of hyperparameters we found, while TRPO was not tuned. Both techniques rely on the same theoretical base: both fall under the policy gradient-based methods of RL, both aim to optimize a surrogate function for the return based on the Performance Difference Lemma (PDL) proposed in (Kakade & Langford, 2002), both aim to do conservative updates to the policy parameters to ensure a stable learning process using the Kullback-Leibler (KL) Divergence between the iterations of the policy's distribution. The main difference lies in removing the use of second-order information in the optimization process (TRPO), replacing it with a clipped version (PPO). Each approach has its advantages and disadvantages. We faced them during the training process, and it also reflected in the results. For instance, PPO faced challenges when trying to reduce

the optimality gap in the vicinity of the solution, and adding more iterations destabilized the training (even using a small learning rate). TRPO suffered when facing a nonconvenient initialization, which is a common problem in second-order methods, limiting its capability of exploiting sampling (more trajectories would lead to a better Hessian estimation) to optimize in fewer iterations with an adaptive learning rate.

Table 3. RL case for **agent i**: detail of the state composition with the units of each feature, action definition, and reward design.

State: $\mathbf{s_{t,i}}$			
Hour	t	$\in \{1, 2, \ldots, 23\}$	$-$
Day Type	d	$\in \{1, 2, \ldots, 7\}$	$-$
Solar Generation	$E_{t,i}^{\text{solar}}$	$\in \mathbb{R}_{\geq 0}$	kWh
Non-Shiftable Load	$E_{t,i}^{\text{load}}$	$\in \mathbb{R}_{\geq 0}$	kWh
Electrical Storage SOC	$B_{t,i}^{\text{soc}}$	$\in [0, 1]$	$\%$
Net Electricity Consumption	$E_{t,i}^{\text{net}}$	$\in \mathbb{R}$	kWh
Predicted Load (4h)	$E_{t,i}^{\text{load+4}}$	$\in \mathbb{R}_{\geq 0}$	kWh
Predicted Load (6h)	$E_{t,i}^{\text{load+6}}$	$\in \mathbb{R}_{\geq 0}$	kWh
Predicted Load (12h)	$E_{t,i}^{\text{load+12}}$	$\in \mathbb{R}_{\geq 0}$	kWh
Predicted Load (24h)	$E_{t,i}^{\text{load+24}}$	$\in \mathbb{R}_{\geq 0}$	kWh
Direct Solar Irradiance	$I_{t,i}^{\text{direct}}$	$\in \mathbb{R}$	W/m^2
Predicted Solar Irradiance (4h)	$I_{t,i}^{\text{direct+4}}$	$\in \mathbb{R}_{\geq 0}$	W/m^2
Predicted Solar Irradiance (6h)	$I_{t,i}^{\text{direct+6}}$	$\in \mathbb{R}_{\geq 0}$	W/m^2
Predicted Solar Irradiance (12h)	$I_{t,i}^{\text{direct+12}}$	$\in \mathbb{R}_{\geq 0}$	W/m^2
Predicted Solar Irradiance (24h)	$I_{t,i}^{\text{direct+24}}$	$\in \mathbb{R}_{\geq 0}$	W/m^2
Electricity Pricing	$R_{t,i}^{\text{buy}}$	$\in \mathbb{R}_{\geq 0}$	USD$/kWh$
Selling Pricing	$R_{t,i}^{\text{sell}}$	$\in \mathbb{R}_{\geq 0}$	USD$/kWh$
Carbon Intensity	$R_{t,i}^{\text{carbon}}$	$\in \mathbb{R}_{\geq 0}$	kgCO$_2/kWh$
Other Relevant Values			
Battery Capacity	B_i^{cap}	$\in \mathbb{R}_{\geq 0}$	kWh
Energy Cost Weight	W_C	$\in [0, 1]$	$-$
Energy Emissions Weight	W_G	$\in [0, 1]$	$-$
Note: $W_C + W_G = 1$			
Action: $\mathbf{a_{t,i}}$			
Battery charge/discharge setpoint	$B_{t,i}^{\text{action}}$	$\in [-1, 1]$	$\%$
Reward: $\mathbf{r_{t,i}}$			
$P_{i,t} = B_i^{\text{cap}} * \begin{cases} \max\left\{0, B_{t-1,i}^{\text{action}} - (1 - B_{t-1,i}^{\text{soc}})\right\} & \text{if } B_{t-1,i}^{\text{action}} \geq 0 \\ \max\left\{0, \lvert B_{t-1,i}^{\text{action}}\rvert - B_{t-1,i}^{\text{soc}}\right\} & \text{o.w.} \end{cases}$			
$C_{i,t} = \max\left\{0, E_{t,i}^{\text{net}} + P_{i,t}\right\} * R_{t,i}^{\text{buy}} + \min\left\{0, E_{t,i}^{\text{net}} + P_{i,t}\right\} * R_{t,i}^{\text{sell}}$			
$G_{i,t} = \max\left\{0, E_{t,i}^{\text{net}} + P_{i,t}\right\} * R_{t,i}^{\text{carbon}}$			
$r_{t,i} = -\left(C_{i,t} * W_C + G_{i,t} * W_G\right)$			