B2RL: An open-source Dataset for Building Batch Reinforcement Learning

Hsin-Yu Liu*
hyl001@eng.ucsd.edu
University of California, San Diego
La Jolla, CA, USA

Xiaohan Fu
x5fu@eng.ucsd.edu
University of California, San Diego
La Jolla, CA, USA

Bharathan Balaji**
bhabalaj@amazon.com
Amazon
USA

Rajesh Gupta
gupta@eng.ucsd.edu
University of California, San Diego
La Jolla, CA, USA

Dezhi Hong**
hondezhi@amazon.com
Amazon
USA

ABSTRACT

Batch reinforcement learning (BRL) is an emerging research area in the RL community. It learns exclusively from static datasets (i.e. replay buffers) without interaction with the environment. In the offline settings, existing replay experiences are used as prior knowledge for BRL models to find the optimal policy. Thus, generating replay buffers is crucial for BRL model benchmark. In our B2RL (Building Batch RL) dataset, we collected real-world data from our building management systems, as well as buffers generated by several behavioral policies in simulation environments. We believe it could help building experts on BRL research. To the best of our knowledge, we are the first to open-source building datasets for the purpose of BRL learning.

ACM Reference Format:

Hsin-Yu Liu*, Xiaohan Fu, Bharathan Balaji**, Rajesh Gupta, and Dezhi Hong**. 2022. B2RL: An open-source Dataset for Building Batch Reinforcement Learning. In *Third ACM SIGEnergy Workshop on Reinforcement Learning for Energy Management in Buildings & Cities (RLEM) (RLEM '22), November 9–10, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3563357.3566164

1 INTRODUCTION

Reinforcement learning (RL) is widely studied in the building research area. Most studies focus on RL learning in an online paradigm [5, 11, 19, 25, 28, 30], assuming there is a simulation environment for RL models to interact with during training and evaluation stages before real-world deployment. Simulators such as EnergyPlus [3] and TRNSYS [14] are used to simulate the thermal states of a building. However, designing and calibrating such models for a large building is time-consuming and requires expertise.

In real-world scenarios, most large buildings are controlled via building management systems (BMS), where thermal data can be stored in database. With advances in sensing technologies and machine learning, data-driven models have been more popular in recent research. Batch reinforcement learning, a data-driven approach that learns only from fixed dataset generated with unknown behavioral policy, has not been explored widely in the building control community. BRL models are capable of learning the optimal policy without accurate environment models or simulation environments as oracles. In our study, we open-source both our dataset (https://github.com/HYDesmondLiu/B2RL) extracted from real building and the one generated with Sinergym [12], a building RL simulation environment which integrates EnegryPlus and BCVTB [26] with OpenAI Gym [2] interface. Furthermore, we experiment with several state-of-the-art BRL methods. The experimental results could be re-used as benchmarks for algorithm comparison.

*Corresponding author.
** Work unrelated to Amazon.

BuildSys '22, November 9–10, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9890-9/22/11.
https://doi.org/10.1145/3563357.3566164

2 RELATED WORK

2.1 Building batch reinforcement learning

Previously, several studies implement fitted Q-iteration (FQI) and batch Q-learning [20, 21, 23, 27]. However, for FQI and batch Q-learning, they are based on pure off-policy algorithms. Fujimoto et al. [23] show that off-policy methods exacerbate the extrapolation error in a pure offline setting. These errors are attributed to Q-network training on historical data but exploratory actions yield policies which are different from the behavioral ones.

Recently, several studies related to building deep BRL research have emerged. Zhang et al. [29] apply CQL [16] on the CityLearn [24] testbed as simulator. Liu et al. [17] incorporates a Kullback-Leibler term in Q-update to penalize policies that are far from the previous one to improve from state-of-the-art BRL algorithm and deploy in real environments without setting up simulators.

2.2 Batch reinforcement learning datasets

To our best knowledge, the only open-source BRL dataset is the D4RL dataset [8]. They have generated various robotic control datasets. In our study, we open-source two building datasets, one contains real building buffers extracted from our building database with sensor readings, setpoints control history, and the estimated energy consumption calculated by Zonepac [1]. Then, we process them as Markov Decision Process (MDP) tuples. The other one is a set of buffers that contain different qualities of transitions generated by pre-trained behavioral agents with simulation environments.

3 APPROACH AND RESULTS

3.1 Real building buffers

3.1.1 Data acquisition. The real building buffer is extracted from the readings of student labs in one of the school buildings. The amount of datapoints in the buffers ranges from 170~260K, depending on the number of rooms involved and missing values. We obtain data of an entire year, from the beginning of July 2017 to the end of June 2018 for 15 rooms across 3 floors. The RL setup in our experiments is listed as below:

- *State*: Indoor air temperature, actual supply airflow, outside air temperature, and humidity.
- *Action*: Zone air temperature setpoint and actual supply airflow setpoint. Both are in continuous space and the action spaces are normalized in the range of $[-1, 1]$ as a standard RL settings.
- *Reward*: Our reward function is a linear combination of thermal comfort and energy consumption. The reward function at time step t is:
$$R_t = -\alpha|TC_t| - \beta P_t, \tag{1}$$
where α, β are the weights balancing different objectives and could be tuned to meet specific goals, TC_t is the thermal comfort index at time t, P_t is the HVAC power consumption at time t. We compute P_t attributed to a thermal zone using heat transfer equations [1].

3.1.2 BRL benchmarks.

- Batch-constrained deep Q-learning (BCQ) [10]: BCQ is a model-free RL method that mitigates extrapolation errors induced by incorrect value estimation of out-of-distribution actions selected out of existing dataset.
- Bootstrapping Error Accumulation Reduction (BEAR) [15]: BEAR identifies bootstrapping error as a key source of BRL instability. The algorithm mitigates out-of-distribution action selection by searching over the set of policies that is akin to the behavioral policy.
- Pessimistic Q-Learning (PQL) [18]: PQL uses pessimistic value estimates in the low-data regions in the Bellman optimality equation as well as the evaluation back-up. It can yield stronger guarantees when the concentrability assumption does not hold. PQL learns from policies that satisfy a bounded density ratio assumption similar to on-policy policy gradient methods.

3.1.3 Experiment details. Each algorithm is run in one room on each floor for an entire week so that outside air temperature (OAT) is the same. For instance, in one week we run algorithm A in rooms in the same stack on different floors, e.g. 2144, 3144, and 4144, and at the same time algorithm B runs on 2146, 3146, and 4146, and so forth. In each room, we train the algorithm for 1,000 time steps, which is about one week. We evaluate each algorithm in three different rooms (one room from each floor: 2F, 3F, and 4F). These rooms are of roughly the same size and occupancy capacity. Each time step is 10 minute due to the data writing rate in our BMS. More details of the experiments are described previously in our previous study [29].

Fig. 2 shows the learning curves of each algorithm, where each solid line is the average reward of all runs for the same method; semi-transparent bands represent the range of all runs for a particular algorithm. And gray dotted vertical lines indicate 00:00AM of each day. The horizontal black dotted line is the average reward in the buffer. Fig. 3 shows the analysis of the optimization objectives in the reward function, for energy consumption, the default control method rule-based contorl (RBC) method is normalized to 1. And for thermal comfort we are showing absolute averaged values.

As we need to calculate the thermal comfort level as required by our reward function, we adopt the widely used predicted mean vote (PMV) [6] measure as our thermal comfort index. In this metric, thermal comfort satisfaction ranges from -3 (cold) to 3 (hot), where PMV within the range of -0.5 to 0.5 is considered as thermal comfortable. We adopt the ASHRAE RP-884 thermal comfort data set [4] and train a simple gradient boosting tree (GBT) model [13] to predict the thermal comfort by taking the current thermal states given by our building system in real-time.

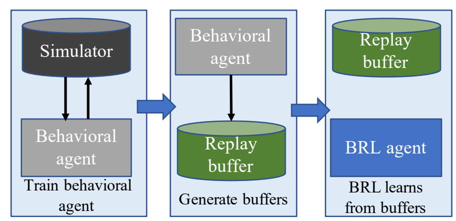

Figure 1: Flow of buffer generation and BRL training

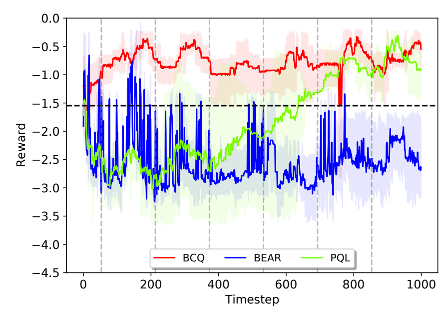

Figure 2: Episode reward comparison in real building

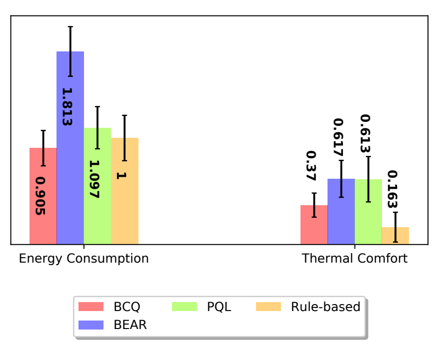

Figure 3: Optimization objectives analysis in real building

3.2 Simulated buffers

3.2.1 Data acquisition. We adopt Sinergym, an open-source simulation and control framework for training RL agents [12]. It is compatible with EnergyPlus models using Python APIs. Our approach follows the BRL paradigm. (1) We first train behavioral RL agents for $500K$ timesteps and select the one that gives the highest average score as the expert agent. Then we run on a 5-zone building, which is a single floor building divided into 5 zones, 1 interior and 4 exterior with 3 weather types: cool, hot, and mixed in continuous settings. We also experiment on two different kinds of response type, deterministic and stochastic. Then we generate expert buffer with $500K$ transitions as the expert buffer. (2) A medium buffer is generated when the behavioral agent is trained "halfway", which means the evaluation score reaches half of the expert agents' final average scores. (3) We randomly initialize the agent, which samples action from allowed action spaces with uniform distribution to generate buffers. (See Fig. 1)

- State: Site outdoor air dry bulb temperature, site outdoor air relative humidity, site wind speed, site wind direction, site diffuse solar radiation rate per area, site direct solar radiation rate per area, zone thermostat heating setpoint temperature, zone thermostat cooling setpoint temperature, zone air temperature, zone thermal comfort mean radiant temperature, zone air relative humidity, zone thermal comfort clothing value, zone thermal comfort Fanger model PPD, zone people occupant count, people air temperature, facility total HVAC electricity demand rate, current day, current month, and current hour.
- Action: Heating setpoint and cooling setpoint in continuous settings.
- Reward: We follow the default linear reward settings, it considers the energy consumption and the absolute difference to temperature comfort.

3.2.2 BRL benchmarks. With various qualities of buffers, we compare several most representative benchmarks in the BRL literature and summarize the average scores and standard deviation in the last 5 evaluations across 3 random seed runs (see Table 1). The scores of random policy is normalized to 0 and expert policy is normalized to 100.

- TD3+BC: An offline version of TD3, it simply adds a behavior cloning term to regularize actor policy towards behavioral policy [9] combined with mini-batch Q-values and buffer states normalization for stability improvement.
- CQL: Conservative Q-learning [16], derived from SAC, learns a lower-bound estimates of the value function, by regularizing the Q-values during training.
- BC: Behavior cloning, we train a VAE to reconstruct action given state. It simply imitate the behavioral agent without reward signals.

We train each algorithm for $500K$ timesteps. For every $25K$ timesteps of training we evaluate the models for one episode. As an example, we illustrate BRL learning curves with expert buffers in Fig. 4.

4 CONCLUSION AND FUTURE WORKS

We open-source our building control datasets for both real buildings and simulation environments for BRL learning. The goal is to encourage building domain experts to explore opportunities in building-BRL research. We provide these datasets for researchers to implement fast prototyping without generating buffers on their own. Recently, many building-RL libraries are published [7, 22, 24] for the purpose of building RL training without the need to set up thermal simulators beforehand. Our future work is to generate more diverse buffers with various building environments and different weather types for BRL benchmarks.

ACKNOWLEDGEMENT

This work was supported in part by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

REFERENCES

[1] Bharathan Balaji, Hidetoshi Teraoka, Rajesh Gupta, and Yuvraj Agarwal. 2013. Zonepac: Zonal power estimation and control via hvac metering and occupant feedback. In *BuildSys*. 1–8.
[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *arXiv preprint arXiv:1606.01540* (2016).
[3] Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. 2001. EnergyPlus: creating a new-generation building energy simulation program. *Energy and buildings* 33, 4 (2001), 319–331.
[4] Richard J De Dear. 1998. A global database of thermal comfort field experiments. *ASHRAE transactions* 104 (1998), 1141.
[5] Xianzhong Ding, Wan Du, and Alberto E Cerpa. 2020. MB2C: Model-Based Deep Reinforcement Learning for Multi-zone Building Control. In *BuildSys*. 50–59.
[6] Povl O Fanger et al. 1970. Thermal comfort. Analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering.* (1970).
[7] Arduin Findeis, Fiodar Kazhamiaka, Scott Jeen, and Srinivasan Keshav. 2022. Beobench: a toolkit for unified access to building simulations for reinforcement learning. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*. 374–382.
[8] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219* (2020).
[9] Scott Fujimoto and Shixiang Shane Gu. 2021. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems* 34 (2021), 20132–20145.
[10] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *ICML*. PMLR, 2052–2062.
[11] Guanyu Gao, Jie Li, and Yonggang Wen. 2020. DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings via Reinforcement Learning. *IEEE Internet of Things Journal* 7, 9 (2020), 8472–8484.
[12] Javier Jiménez-Raboso, Alejandro Campoy-Nieves, Antonio Manjavacas-Lucas, Juan Gómez-Romero, and Miguel Molina-Solana. 2021. Sinergym: a building simulation and control framework for training reinforcement learning agents. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 319–323.
[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *NIPS* 30 (2017), 3146–3154.
[14] SA Klein. 1976. University of Wisconsin-Madison Solar Energy Laboratory. *TRNSYS: A transient simulation program. Eng. Experiment Station* (1976).
[15] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949* (2019).
[16] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 1179–1191.
[17] Hsin-Yu Liu, Bharathan Balaji, Sicun Gao, Rajesh Gupta, and Dezhi Hong. 2022. Safe HVAC Control via Batch Reinforcement Learning. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 181–192.

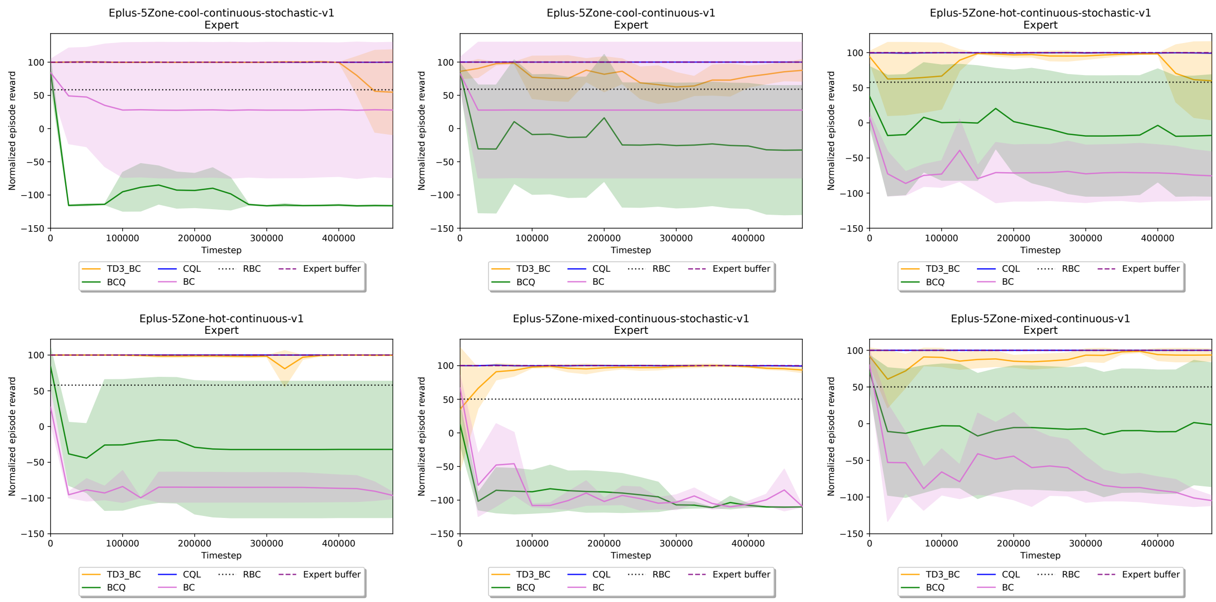

Figure 4: Learning curves of BRL models that learn from expert buffers. Solid line shows the averaged value across three random seeds per algorithm, and the half-transparent region indicates the range with one standard deviation.

Environment	Buffer	TD3+BC	CQL	BCQ	BC
hot-deterministic	Expert	**99.72±0.1**	**100.00±0.00**	-32.02±0.07	-89.2±3.95
hot-deterministic	Medium	-49.59±8.19	**67.65±17.06**	13.41±16.59	-12.55±7.27
hot-deterministic	Random	-45.73±15.13	-23.19±4.52	**69.21±18.52**	-26.74±15.91
mixed-deterministic	Expert	**94.67±2.04**	**100.00±0.00**	-6.22±5.24	-95.46±6.6
mixed-deterministic	Medium	36.23±4.31	37.36±19.31	**64.46±0.65**	-103.4±2.12
mixed-deterministic	Random	-13.72±22.25	-23.46±20.33	-65.30±20.40	-27.82±11.79
cool-deterministic	Expert	81.11±5.24	**100.00±0.00**	-29.75±3.18	27.76±0
cool-deterministic	Medium	-49.97±0.00	55.44±6.46	**70.19±17.06**	10.48±22.11
cool-deterministic	Random	-58.40±3.21	12.99±2.28	**27.77±31.39**	8.62±41.97
hot-stochastic	Expert	77.69±17.18	**99.49±0.20**	-15.35±5.92	-72.86±1.73
hot-stochastic	Medium	-14.85±0.00	**39.93±2.64**	-62.21±19.31	-10.45±12.85
hot-stochastic	Random	-1.82±2.68	**36.65±11.95**	-1.24±14.80	31.22±13.51
mixed-stochastic	Expert	**96.61±2.13**	**99.77±0.26**	-108.38±2.58	-102.02±9.32
mixed-stochastic	Medium	9.49±0.00	**80.13±8.19**	70.75±6.46	-107.41±3.41
mixed-stochastic	Random	28.02±8.69	**94.05±2.08**	-109.47±0.17	38.66±24.64
cool-stochastic	Expert	78.27±20.01	**99.97±0.12**	-115.86±0.41	28.15±0.35
cool-stochastic	Medium	16.09±0.00	**81.57±4.31**	-11.55±2.64	-50.37±2.45
cool-stochastic	Random	-44.33±16.01	-97.35±2.09	-53.92±10.07	**25.44±13.42**
Sum		339.50±127.23	**960.99±101.81**	-295.49±175.47	-527.93±193.48

Table 1: Average normalized score over the final 5 evaluations and 3 random seeds. ± corresponds to standard deviation over the last 5 evaluations across runs.

[18] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. 2020. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202* (2020).

[19] Naren Srivaths Raman, Adithya M Devraj, Prabir Barooah, and Sean P Meyn. 2020. Reinforcement learning for control of building HVAC systems. In *2020 American Control Conference (ACC)*. IEEE, 2326–2332.

[20] Frederik Ruelens, Bert J Claessens, Stijn Vandael, Bart De Schutter, Robert Babuška, and Ronnie Belmans. 2016. Residential demand response of thermostatically controlled loads using batch reinforcement learning. *IEEE Transactions on Smart Grid* 8, 5 (2016), 2149–2159.

[21] Frederik Ruelens, Bert J Claessens, Stijn Vandael, Sandro Iacovella, Pieter Vingerhoets, and Ronnie Belmans. 2014. Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning. In *2014 Power Systems Computation Conference*. IEEE, 1–7.

[22] Paul Scharnhorst, Baptiste Schubnel, Carlos Fernández Bandera, Jaume Salom, Paolo Taddeo, Max Boegli, Tomasz Gorecki, Yves Stauffer, Antonis Peppas, and Chrysa Politi. 2021. Energym: A building model library for controller benchmarking. *Applied Sciences* 11, 8 (2021), 3518.

[23] José Vázquez-Canteli, Jérôme Kämpf, and Zoltán Nagy. 2017. Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration. *Energy Procedia* 122 (2017), 415–420.

[24] José R Vázquez-Canteli, Sourav Dey, Gregor Henze, and Zoltán Nagy. 2020. CityLearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management. *arXiv preprint arXiv:2012.10504* (2020).

[25] Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (2020), 115036.

[26] Michael Wetter, Philip Haves, and Brian Coffey. 2008. *Building controls virtual test bed*. Technical Report. Lawrence Berkeley National Laboratory.

[27] Lei Yang, Zoltan Nagy, Philippe Goffin, and Arno Schlueter. 2015. Reinforcement learning for optimal control of low exergy buildings. *Applied Energy* 156 (2015), 577–586.

[28] Chi Zhang, Sanmukh R Kuppannagari, Rajgopal Kannan, and Viktor K Prasanna. 2019. Building HVAC scheduling using reinforcement learning via neural network based model approximation. In *BuildSys*. 287–296.

[29] Chi Zhang, Sanmukh Rao Kuppannagari, and Viktor K Prasanna. 2022. Safe Building HVAC Control via Batch Reinforcement Learning. *IEEE Transactions on Sustainable Computing* (2022).

[30] Zhiang Zhang and Khee Poh Lam. 2018. Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. In *BuildSys*.