**Title**

Deep Reinforcement Learning in Buildings



**Authors**

Prakash, Anand Krishnan
Touzani, Samir
Kiran, Mariam
et al.

**Publication Date**

2020-11-17

**DOI**

10.1145/3427773.3427868

Peer reviewed

DOI 10.1145/3427773.3427868

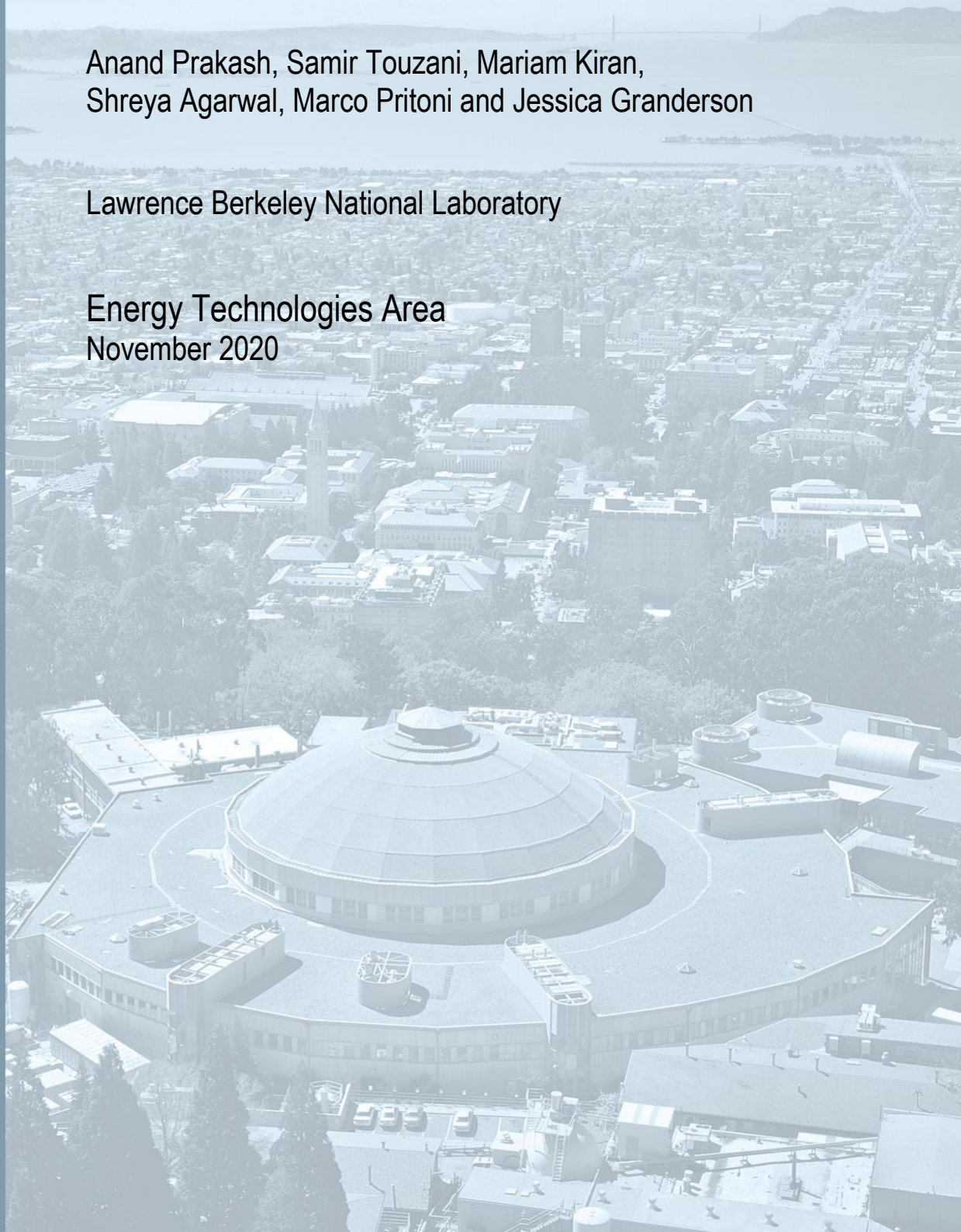
Lawrence Berkeley National Laboratory

# Deep Reinforcement Learning in Buildings: Implicit Assumptions and their Impact

Anand Prakash, Samir Touzani, Mariam Kiran,
Shreya Agarwal, Marco Pritoni and Jessica Granderson

Lawrence Berkeley National Laboratory

Energy Technologies Area
November 2020



# Deep Reinforcement Learning in Buildings: Implicit Assumptions and their Impact

Anand Krishnan Prakash
Lawrence Berkeley National Laboratory
akprakash@lbl.gov

Samir Touzani
Lawrence Berkeley National Laboratory
stouzani@lbl.gov

Mariam Kiran
Lawrence Berkeley National Laboratory
mkiran@es.net

Shreya Agarwal
Lawrence Berkeley National Laboratory
shreyaagarwal@lbl.gov

Marco Pritoni
Lawrence Berkeley National Laboratory
mpritoni@lbl.gov

Jessica Granderson
Lawrence Berkeley National Laboratory
jgranderson@lbl.gov

## ABSTRACT

As deep reinforcement learning (DRL) continues to gain interest in the smart building research community, there is a transition from simulation-based evaluations to deploying DRL control strategies in actual buildings. While the efficacy of a solution could depend on a particular implementation, there are common obstacles that developers have to overcome to deliver an effective controller. Additionally, a deployment in a physical building can invalidate some of the assumptions made during the controller development. Assumptions on the sensor placement or on the equipment behavior can quickly come undone. This paper presents some of the significant assumptions made during the development of DRL based controllers that could affect their operations in a physical building. Furthermore, a preliminary evaluation revealed that controllers developed with some of these assumptions can incur twice the expected costs when they are deployed in a building.

## CCS CONCEPTS

• **Computing methodologies** → **Reinforcement learning**; **Learning from demonstrations**; *Control methods*; • **Computer systems organization** → **Sensors and actuators**.

## KEYWORDS

reinforcement learning, smart buildings, control systems, distributed energy resources

**ACM Reference Format:**
Anand Krishnan Prakash, Samir Touzani, Mariam Kiran, Shreya Agarwal, Marco Pritoni, and Jessica Granderson. 2020. Deep Reinforcement Learning in Buildings: Implicit Assumptions and their Impact. In *The 1st International Workshop on Reinforcement Learning for Energy Management in Buildings & Cities (RLEM'20), November 17, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3427773.3427868


*RLEM'20, November 17, 2020, Virtual Event, Japan*

ACM ISBN 978-1-4503-8193-2/20/11.
https://doi.org/10.1145/3427773.3427868

## 1 INTRODUCTION

As researchers start to look beyond model predictive control (MPC) to optimize the operation of building equipment, reinforcement learning (RL) has been gaining traction. With advances in deep learning, emerging infrastructure are increasingly using data driven artificial intelligence based strategies to control, manage and also optimize their usage [3]. Using deep neural networks to handle the high dimensionality of complex systems like buildings, Deep Reinforcement Learning based control algorithms (*DRL controllers*) can help in optimizing the operations of these systems. The 'model-free' nature of DRL and the adaptability to handle unknown conditions are unique selling points that these solutions offer over MPC [7]. The term 'model-free', refers to a controller that uses a data-driven approach to learn the optimal actions for each given state instead of understanding the first principles involved in the transition between the actions and new state [6]. Preliminary deployments in residential and small commercial buildings have already demonstrated improvements in energy savings and in the indoor environmental quality [7]. However, developing DRL is not without its challenges and Wang and Hong (2020) [7] have discussed those in detail. They are further compounded when the objective is to deploy such a controller in an actual building.

This paper focuses on the *implicit assumptions* made during the development of a DRL controller on the training processes and on the building systems. For example, while the developers of controllers usually account for the possibility of faulty equipment in buildings, the adequacy of the sequence of operations programmed in these equipment are often not questioned. Furthermore, developers expect the model-free nature of DRL based controllers to steer through such discrepancies with ease. The main contributions are:

- A description of the significant implicit assumptions made during the development of a DRL controller that no longer hold true when it is deployed in a building.
- An evaluation of the impact of the assumptions on the actions generated by a DRL controller aiming to optimize the operations of the equipment in a single zone building.

## 2 METHODOLOGY

Developing a DRL controller involves formulating an agent that is in a partially observable environment and learning the best decisions through interacting with the environment. The agent observes

environment snapshots, and chooses an action, receiving a reward value for this action in the current state. It continues to receive feedback on its actions until a terminal state is reached. The objective is to maximize the cumulative reward over all actions in the time the agent is active [5].

Nevertheless, permitting a DRL controller to train directly in an actual building (*on-policy* learning) could result in discomfort to the occupants and damage to the equipment. Hence, the agent is trained using data from the building (*off-policy* learning). However, as operational data from an existing building might be insufficient to represent all possible system states, oftentimes simulation models of the building (using MATLAB, EnergyPlus etc.) are used to train the DRL controller [7].

The authors of this paper have used DRL controllers for the supervisory control of a heating, ventilation and air conditioning (HVAC) system and a battery storage in a single zone building with on-site solar generation. The objective of the controller is to minimize the total energy costs incurred while maintaining indoor thermal comfort. More specifically, Deep Deterministic Policy Gradient (DDPG) [2] has been used for developing the controller. DDPG is an actor-critic DRL algorithm that can be used in continuous action space. A training framework that uses a physics-based simulation component was implemented to learn end-to-end control policies. Energyplus [4] was used to model the building and HVAC system while Modelica [8] was used to simulate the battery and PV.

The following sections describe some of the significant assumptions made during the development phase and studies their impact during operations, with the intention that the analyses conducted in this paper can help other DRL practitioners who might've made similar assumptions, accelerate their research and deployment studies.

## 3 IMPLICIT ASSUMPTIONS

This section discusses three assumptions made by the authors during the development of the DRL controller, related to: 1) the random seed used during training 2) data used for the training and 3) equipment behavior in an existing building.

### 3.1 Impact of the Random Seed

During training of a DRL controller, a seed value is used to randomize the initialization and to ensure reproducibility. Two controllers starting with the same seed would produce the identical actions at each step and conversely, two controllers trained with different seeds would generate slightly different actions at each state. However, it was assumed that with extensive training, the variance of the actions generated by differently-seeded controllers could be minimized and hence choosing the seed value (often done at random) had little repercussions. This was quickly disproved when the authors identified a controller configuration that was able to generate optimal actions after training using `seed1`, but was unable to even converge when trained using `seed2`. Figure 1 shows the distribution of reward values received at each step by four similarly configured DRL controllers trained using different seeds. Even though the median values of these distributions are quite similar, the different heights of the box plots point to the different actions generated.

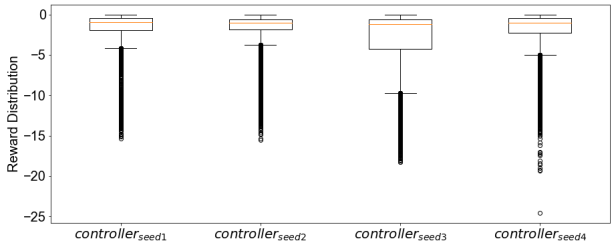



**Figure 1: Varying reward distribution of similarly configured DRL controllers trained with different seeds indicating different actions**

### 3.2 Training Data

Due to the adaptability of DRL controllers to new system states and the possibility of online learning during the actual operations, developers tend to assign less emphasis on the simulated building representation and on the quality of the training data. Figure 2 illustrates two streams of power consumed by the supply air fan of an air handler unit (AHU): one from a calibrated energy model and the other from an uncalibrated energy model. Even though it was obvious that training a DRL controller on a stream of data from an uncalibrated source against a calibrated source would produce different controllers, it was assumed that when the real system state was fed as input, the actions generated by both the trained controllers would be quite similar. This belief was reinforced by the fact that the power consumed by the supply air fan was only a small fraction of the total building load (median ratio of 8.52%).

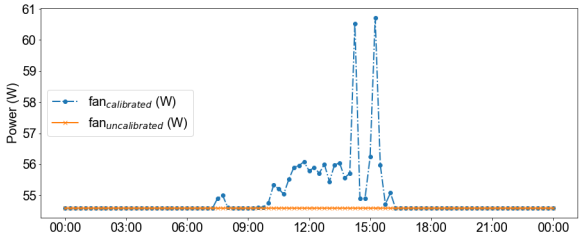



**Figure 2: Different power consumption of supply air fan (`fan`$_{\texttt{calibrated}}$, `fan`$_{\texttt{uncalibrated}}$) used to train DRL controllers**

### 3.3 Equipment Behavior

As the DRL controller was trained through interactions with a simulated representation of the building, it was implicitly assumed that an actuator would always meet the generated setpoint (or the action). However, in a physical building, this is usually not the case. Even without the setpoints being out of bounds or a network interruption causing the control signal to be dropped, the physical limitations or the capacity constraints on the building's control system could prevent it from successfully conditioning the environment to meet its setpoint. Figure 3 illustrates an instance of this assumption failing for an AHU's supply air temperature when despite the obvious lag in time, the actual values are different from the requested setpoints.

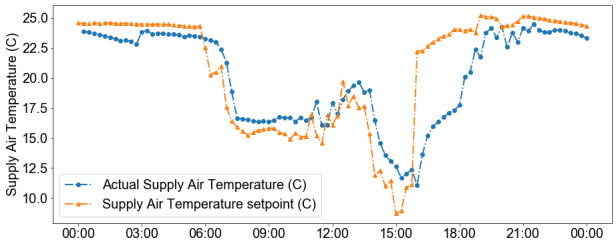



**Figure 3: DRL generated supply air temperature setpoint v/s the actual supply air temperature of an AHU**

## 4 EVALUATION

The evaluation is conducted by deploying the DRL controllers in an existing building (described below). The controller queries the state of the system periodically (configured to run every fifteen minutes) and generates certain actions. Based on the the objective of each experiment, the controllers might be set up to operate in one of the following modes:

- *shadow-mode*: While DRL generates setpoints based on the actual state of the system, they are not applied to the actual building equipment; an independent sequence of operation would be controlling the equipment. This mode allows the comparison of setpoints generated by multiple controllers at the same system state.
- *closed-loop-mode*: Here the setpoints generated by the DRL controller are actually set on the controller to trigger a change in operation.

### 4.1 Experimental Set up

DRL controllers based on the DDPG algorithm have been used in this evaluation. They have been deployed in a single zone building in FLEXLAB [1], an experimental building facility at Lawrence Berkeley National Laboratory, with the objective to minimize energy costs incurred while maintaining a comfortable zone air temperature band of 21C-24C during occupied hours (7AM-7PM). The building is assumed to be enrolled in a time-of-use (TOU) electricity tariff, with peak prices from 4PM-9PM ($1.20/kWh more than a base price of $0.13/kWh). It is conditioned by an AHU, with on-site solar generation and battery storage to support the load. Correspondingly, during each run (every fifteen minutes) the DRL controllers generates three actions: 1) supply air temperature setpoint for the AHU 2) supply air flow rate setpoint for the AHU and 3) charge/discharge rate setpoint for the battery.

### 4.2 Experiments

Leveraging the test setup mentioned above, we conduct the following experiments to study the impact of the assumptions discussed in Section 3.

*4.2.1 Different seeds.* As shown in Figure 1, changing the initial seed values used in the training of the DRL controller does produce different controllers. The impact of using different seeds was quantified by comparing the setpoints generated by a DRL controller trained using seed1 ($\texttt{controller}_{\texttt{seed1}}$) against those generated by a controller trained using seed2 ($\texttt{controller}_{\texttt{seed2}}$). This comparison was conducted in *shadow-mode* and Figure 4 visualizes the supply air temperature setpoints generated by both the controllers.

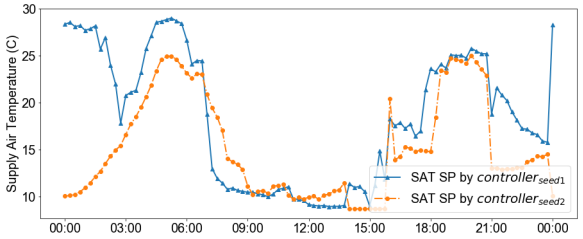



**Figure 4: AHU supply air temperature (SAT) setpoints generated by two DRL controllers with the same configuration, but trained using different seed values**

*4.2.2 Quality of training data.* By comparing the setpoints generated by two similar DRL controllers, trained on different datasets ($\texttt{controller}_{\texttt{calibrated}}$, $\texttt{controller}_{\texttt{uncalibrated}}$), both running in *shadow-mode*, the impact of training data on a controller's actions were assessed (illustrated in Figure 5). The training data only differ in the power consumption of the supply air fan power and correspond to the $\texttt{fan}_{\texttt{calibrated}}$ and $\texttt{fan}_{\texttt{uncalibrated}}$ timeseries illustrated in Figure 2.

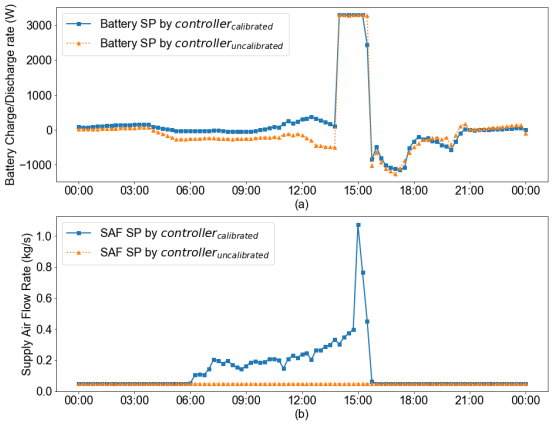



**Figure 5: (a) Battery charge/discharge rate and (b) AHU supply air flow rate (SAF) setpoints generated by two DRL controllers with the same configuration, but trained on slightly different datasets**

*4.2.3 Meeting the setpoints.* Supervisory DRL controllers affects the building environment by generating setpoints for the equipment that condition this environment. However, Figure 3 provides evidence that equipment do not always reach their setpoints. In order to study the impact of inability of building equipment to meet the target setpoint on DRL, a simulated battery model ($\texttt{battery}_{\texttt{model}}$) with the same specifications as the real battery ($\texttt{battery}_{\texttt{real}}$) was constructed, with the key difference being the ability of the $\texttt{battery}_{\texttt{model}}$ to charge and discharge based on the

**Table 1: A table of Mean Absolute Error (MAE) between setpoints generated by the DRL set up under test and those generated by the baseline DRL set up for each experiment**

| Experiment | Test Controller | Baseline Controller | Affected Setpoint | MAE |
|---|---|---|---|---|
| 1. Different seeds | a DRL controller trained on $\texttt{seed}_2$ | a DRL controller trained on $\texttt{seed}_1$ | supply air temperature | 4.28C |
| | | | supply air flow rate | 0.06kg/s |
| | | | battery charge/discharge rate | 162.67W |
| 2. Quality of training data | a DRL controller trained on partly-calibrated data | a DRL controller trained on well-calibrated data | supply air temperature | 3.87C |
| | | | supply air flow rate | 0.082kg/s |
| | | | battery charge/discharge rate | 188.60W |
| 3. Meeting the setpoints | a DRL controller that controls $\texttt{battery}_{\texttt{model}}$ | a DRL controller that controls $\texttt{battery}_{\texttt{real}}$ | supply air temperature | 0.72C |
| | | | supply air flow rate | 0.005kg/s |
| | | | battery charge/discharge rate | 169.62W |

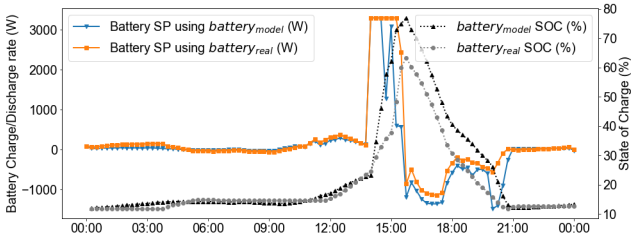



**Figure 6: The different SOC timeseries from $\texttt{battery}_{\texttt{model}}$ and $\texttt{battery}_{\texttt{real}}$ and the DRL generated charge/discharge rate setpoints for the respective batteries**

exact setpoint generated by the DRL controller. This characteristic, in turn creates a new state of charge (SOC) timeseries from the $\texttt{battery}_{\texttt{model}}$. For the evaluation, the setpoints generated by the a DRL controller attached to this $\texttt{battery}_{\texttt{model}}$ are compared to those generated by the same DRL controller attached to the $\texttt{battery}_{\texttt{real}}$, both running in *closed-loop-mode* (as shown in Figure 6)).

## 4.3 Results and Discussion

Table 1 summarizes the results of all the experiments and quantifies the impact of different assumptions on the actions generated by a DRL controller. It can be seen that the largest impact across all three setpoints was incurred when the DRL controller had been trained on a dataset containing uncalibrated power data, even though the uncalibrated part was only a small fraction of the total building load. The minimum supply air fan flow rate throughout the day (Figure 5(b)) also indicates a lack of actuation, resulting in inefficient operations. Table 1 also provides evidence that in spite of training using four years of weather and simulated building operational data, controllers trained using different seeds behave differently. Hence, a controller warrants multiple training cycles, at least to ensure consistent behavior.

*Meeting the setpoints* experiment provides evidence of how unrealistic assumptions made on the behavior of physical equipment could affect a controller. While the battery management was the only noticeable difference, the controller using $\texttt{battery}_{\texttt{real}}$ incurred more than double the energy costs over the course of 24 hours than the controller using the ideal $\texttt{battery}_{\texttt{model}}$. Hence, by modifying the training environments with adjusted rewards function and better state estimations to incorporate such behavior of physical equipment, better controllers can be developed.

## 5 CONCLUSION

While the assumptions presented in this paper are neither algorithm independent nor a complete and exhaustive list, the paper shows that assumptions made during development can significantly impact the performance of a DRL controller when it is deployed in a physical building. Next steps in this work involves studying the impacts of these assumptions on different DRL algorithms and on identifying metrics to quantify the impact on the performance, irrespective of the algorithm or the problem objective. The authors are also looking at developing a better training environment (simulation, reward function etc.) that can incorporate and reflect the negative impacts of these assumptions.

## ACKNOWLEDGMENTS

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors would also like to thank Cedar Blazek, Heather Goetsch and the FLEXLAB team for their support.

## REFERENCES

[1] Berkeley Lab Energy Technologies Area. 2020. *FLEXLAB: Advanced Integrated Building & Grid Technologies Testbed.* https://flexlab.lbl.gov/
[2] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
[3] Volodymyr Mnih and et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533. https://doi.org/10.1038/nature14236
[4] Thierry Nouidui, Michael Wetter, and Wangda Zuo. 2014. Functional mock-up unit for co-simulation import in EnergyPlus. *Journal of Building Performance Simulation* 7, 3 (2014), 192–202.
[5] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* A Bradford Book, USA.
[6] José R. Vázquez-Canteli and Zoltán Nagy. 2019. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy* 235 (2019), 1072 – 1089. https://doi.org/10.1016/j.apenergy.2018.11.002
[7] Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (2020), 115036. https://doi.org/10.1016/j.apenergy.2020.115036
[8] Michael Wetter, Wangda Zuo, Thierry S Nouidui, and Xiufeng Pang. 2014. Modelica buildings library. *Journal of Building Performance Simulation* 7, 4 (2014), 253–270.