Grid-Interactive Multi-Zone Building Control Using Reinforcement Learning with Global-Local Policy Search

Preprint

Xiangyu Zhang, Rohit Chintala, Andrey Bernstein, Peter Graf and Xin Jin

National Renewable Energy Laboratory

Presented at the 2021 American Control Conference
May 25-28, 2021

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-2C00-78000
April 2021

Grid-Interactive Multi-Zone Building Control Using Reinforcement Learning with Global-Local Policy Search

Preprint

Xiangyu Zhang, Rohit Chintala, Andrey Bernstein, Peter Graf and Xin Jin

Suggested Citation
Xiangyu Zhang, Rohit Chintala, Andrey Bernstein, Peter Graf and Xin Jin. 2021. *Grid-Interactive Multi-Zone Building Control Using Reinforcement Learning with Global-Local Policy Search: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-2C00-78000. https://www.nrel.gov/docs/fy21osti/78000.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-2C00-78000
April 2021

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Grid-Interactive Multi-Zone Building Control Using Reinforcement Learning with Global-Local Policy Search

Xiangyu Zhang, Rohit Chintala, Andrey Bernstein, Peter Graf and Xin Jin

***Abstract*— In this paper, we develop a grid-interactive multi-zone building controller based on a deep reinforcement learning (RL) approach. The controller is designed to facilitate building operation during normal conditions and demand response events, while ensuring occupants comfort and energy efficiency. We leverage a continuous action space RL formulation, and devise a two-stage global-local RL training framework. In the first stage, a global fast policy search is performed using a gradient-free RL algorithm. In the second stage, a local fine-tuning is conducted using a policy gradient method. In contrast to the state-of-the-art model predictive control (MPC) approach, the proposed RL controller does not require complex computation during real-time operation and can adapt to non-linear building models. We illustrate the controller performance numerically using a five-zone commercial building.**

I. Introduction

In the U.S., buildings consumed around 39% of the total national energy usage in 2019 (21% for residential and 18% for commercial) [1] and over 70% of the total electricity usage [2]. Specifically, in the commercial sector, heating, ventilation and air-conditioning (HVAC) systems account for 32.7% of the total electricity consumption [3]. Due to its strong coupling with the building thermal dynamics, HVAC control is complicated and is of interest in this and many other studies. Proper HVAC control will not only improve thermal comfort and energy efficiency, but also enable buildings to be grid-interactive, providing valuable demand-side resources in the smart grid paradigm.

Currently, in the literature, model predictive control (MPC) is the mainstream approach for optimal building control. For instance, it has been used for the building cooling system control and demonstrated cost reduction and efficiency improvement in a real-world application [4]. Moreover, for larger buildings with relatively complex HVAC systems, a distributed version of MPC is proposed in [5] for better scalability and more manageable computational burden. See review paper [6] for more examples. Despite the fact that MPC is popular in academic research, its applications are not well-developed in the real world; this is due to two major drawbacks associated with MPC: 1) *high implementation cost*: the on-demand computation of MPC requires optimization problems to be solved repeatedly within each prediction interval, which poses high requirements on the computing hardware and the optimization software; 2) *high modeling cost*: MPC requires building dynamics models, which are building-specific and typically can only be developed by a professional team. The modeling efforts include accurately modeling the building and properly simplifying (e.g., linearizing) it so that the optimization problem can be solved efficiently. As a result, building owners' interest in implementing optimal building control wanes due to these high costs.

In recent years, with the development of artificial intelligence, reinforcement learning (RL) stands out for its excellent performance in optimal decision-making [7]. In building control, RL is also being investigated as an alternative to MPC due to the following merits: 1) The RL control policy can be trained offline and only computationally-inexpensive policy evaluation is conducted during real-time control. 2) Compared with optimization, RL can handle non-linearity and stochasticity more easily. 3) The RL controller can be re-trained periodically to cope with model drift (i.e., change of environment/season or occupants behavior). As a result, many studies now investigate the feasibility of using RL controllers for building control. Deep Q-network (DQN) is applied to implement variable air-flow volume (VAV) control and observe effective cost reduction when compared with a rule-based controller [8]. Similarly, the asynchronous advantage actor-critic (A3C) algorithm is utilized in [9] to optimally control the supply water set point in a building's heating system. In contrast to prior works where a discrete action space is used, an RL algorithm called Zap Q-learning is applied in HVAC control considering a continuous control space [10]. In addition to the building-centric control objectives, using RL to coordinate multiple HVAC units to achieve a load reduction goal during demand response (DR) events while maintaining thermal comfort is investigated in [11].

Though existing RL-based methods have proven effective, we have identified the following knowledge gaps/fields of improvement that are not now addressed or are only partially addressed:

1) Continuous action space should be considered for a more accurate control and to avoid action discretization, which may impact control performance.

The authors are with the U.S. National Renewable Energy Laboratory, Golden, CO 80401, USA. andrey.bernstein@nrel.gov

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

This research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

2) In addition to normal building-centric operation, grid-interactive building control should be encouraged in the future smart grid/smart city paradigm.
3) Zone control should be investigated to enable coordination among zones, as most commercial buildings have multiple thermal zones.
4) To facilitate practical implementation, the perfect forecast of exogenous data, or other hard to acquire information, should not be included as controller inputs.

In this paper, we study a practical RL controller for grid-interactive multi-zone buildings, with the above-mentioned knowledge gaps addressed. Compared to prior works, the continuous policy search space, more complicated controlled system/objective (i.e., multi-zone building and grid service), and the absent of perfect forecast of exogenous information, inevitably pose more difficulties for the RL controller's training. To overcome these challenges, we propose a *two-stage global-local* RL policy search framework. In the first stage, a gradient-free evolution strategy RL (ES-RL) algorithm [12] is leveraged for fast global policy convergence. In the second stage, the proximal policy optimization (PPO) algorithm [13] is utilized for local policy fine-tuning. We illustrate the controller performance numerically using a five-zone commercial building. In particular, we demonstrate that our two-stage approach outperforms the policies trained using ES-RL or PPO alone, and we compare the control performance with an equivalent baseline MPC in both linear and non-linear implementations.

II. Problem Formulation

Buildings are usually divided into several zones, and each zone has its relatively independent thermal dynamics. In a multi-zone commercial building (thermal zone $i \in \mathcal{N} = \{1, 2, ..., N\}$), which utilizes a centralized HVAC system for cooling, the HVAC chiller discharge air temperature (T^{da}) and the cooling air flow rate to each zone ($\dot{m}^i, i \in \mathcal{N}$), can be controlled to provide a comfortable indoor environment for occupants. In this paper, we propose an RL controller that can achieve multi-objective control over the control horizon ($t \in \mathcal{T}$): 1) minimizing the HVAC system's energy consumption; 2) maintaining the indoor comfort; and 3) limiting the maximum power consumption below a certain time-variable threshold $\overline{P}_t$. These objectives are sometimes mutually conflicting, and thus designing a good controller is challenging.

A. HVAC Optimal Control Formulation

Mathematically, the multi-objective optimal HVAC control can be formulated as follows.

First, the thermal discomfort cost measures the temperature deviation from the predefined comfort band $[\underline{T}_t^i, \overline{T}_t^i]$, which is assumed to increase first linearly and then quadratically:

$$\mathcal{D}(T_t^i) := \begin{cases} \max(T_t^i - \overline{T}_t^i, (T_t^i - \overline{T}_t^i)^2), & T_t^i > \overline{T}_t^i \\ \max(\underline{T}_t^i - T_t^i, (\underline{T}_t^i - T_t^i)^2), & T_t^i < \underline{T}_t^i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where T_t^i is the indoor temperature of zone i at step t.

Second, the relationship between the HVAC power consumption (P_t) and control variables (i.e., $\dot{\mathbf{m}}_t = [\dot{m}_t^1, \dot{m}_t^2, ..., \dot{m}_t^N]$ and T_t^{da}) is given by:

$$P(\dot{\mathbf{m}}_t, T_t^{da}) = a(T_t^{out} - T_t^{da}) \sum_{i \in \mathcal{N}} \dot{m}_t^i + b \left(\sum_{i \in \mathcal{N}} \dot{m}_t^i \right)^3 + c \tag{2}$$

in which the first term describes the cooling chiller power (T_t^{out} represents outdoor temperature) and the rest depict the VAV fan power. For each control interval (Δt), the energy consumption is given by:

$$\mathcal{E}_t = P(\dot{\mathbf{m}}_t, T_t^{da}) \Delta t. \tag{3}$$

Third, to account for DR events, a power limit violation penalty term is given by:

$$\mathcal{V}_t := \begin{cases} (P(\dot{\mathbf{m}}_t, T_t^{da}) - \overline{P}_t)^2, & P(\dot{\mathbf{m}}_t, T_t^{da}) \geq \overline{P}_t \\ 0, & \text{otherwise} \end{cases} . \tag{4}$$

During a DR event ($t \in \mathcal{T}_{DR}$), $\overline{P}_t$ is reduced dramatically and the controller needs to reduce P_t accordingly in order to participate in the DR event.

Combining all three objectives together, the optimal control problem is formulated as:

$$\begin{aligned} \underset{\dot{\mathbf{m}}_t, T_t^{da}}{\text{minimize}} \quad & \sum_{t \in \mathcal{T}} \mathbf{w}_t [\kappa_1 \sum_{i \in \mathcal{N}} \mathcal{D}(T_t^i), \kappa_2 \mathcal{E}_t, \kappa_3 \mathcal{V}_t]^\top \\ \text{subject to} \quad & \underline{T}^{da} \leq T_t^{da} \leq \overline{T}^{da}, (\forall t) \\ & \underline{\dot{m}}^i \leq \dot{m}_t^i \leq \overline{\dot{m}}^i, (\forall t, \forall i) \\ & \mathbf{T}_{t+1} = \mathcal{F}\left(\mathbf{T}_t, \dot{\mathbf{m}}_t, T_t^{da}, \mathbf{\Psi}_t\right) (\forall t) . \end{aligned} \tag{5}$$

In (5), κ_i are factors that monetize the three control objectives; $\mathbf{w}_t = [w_t^1, w_t^2, w_t^3]$ is a weighting vector such that $\sum_{i=1}^3 w_t^i = 1$; and $\mathcal{F}$ is the building thermal dynamics model represented by an equality constraint, in which $\mathbf{T}_t \in \mathbb{R}^N$ is the temperature of all zones at step t and $\mathbf{\Psi}_t$ represents all non-controllable exogenous inputs. The details of the building model are discussed next.

B. Building Thermal Dynamics Model

In contrast to optimization-based approaches, RL does not place constraints on the format of $\mathcal{F}$, which can be the actual building, a high-fidelity building simulator (i.e., EnergyPlus [14]) or a reduced-order model (ROM). Considering the low learning efficiency and potentially dangerous exploring actions when learning on the actual building, and the high modeling cost and poor scalability of using EnergyPlus, a ROM, as a data-driven model that represents the middle ground of model accuracy, modeling cost, and scalability, is leveraged in this paper.

A linear parametric modeling approach [15] is used to develop a temperature-predicting model for Zone i using an Auto-Regressive models with eXogenous variables (ARX), which is given by:

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

TABLE I

LIST OF INPUTS CONSIDERED TO PREDICT TEMPERATURE OF ZONE i

Variable	Symbol
Outdoor air temperature	T_t^{out}
Delivered cooling	$Q_t^{HVAC,i}$
Solar heat gain	$Q_t^{solar,i}$
Internal heat gain (reflects occupancy level)	$Q_t^{int,i}$
Other zone's temperature	$T_{t-1}^j, j \in \mathcal{N}, j \neq i$

$$T_t^i = \sum_{j=1}^{n_a} a^{ij} T_{t-j}^i + \sum_{j=1}^{n_b} \mathbf{b}^{ij\top} \mathbf{u}_{t-j}^i \tag{6}$$

where $\mathbf{u}_t^i$ is the input vector, and $a^{ij}, j \in \{1, ..., n_a\}$ and $\mathbf{b}^{ij}, j \in \{1, ..., n_b\}$ are model parameters to be determined. Features to be considered to formulate $\mathbf{u}_t^i$ are shown in Table I, where $Q_t^{solar,i}$ and $Q_t^{int,i}$ are given as exogenous inputs and $Q_t^{HVAC,i}$ is given by:

$$Q_t^{HVAC,i} = C_p \dot{m}_t^i (-T_t^{da} + T_t^i), \tag{7}$$

in which $C_p = 1.0\ kWh/(kg \cdot K)$ representing the specific heat capacity of air. These features are down-selected via feature selection and then the final components of $\mathbf{u}_t^i$ is determined. Note by considering temperatures from other zones, the interactions between different zones are taken into account. Different hyper-parameters are also tested, and in this paper, we use $n_a = 1$ and $n_b = 1$, which can already predict temperature accurately, according to a comparison between outputs of this model and EnergyPlus. In addition, though (6) is linear, the $\dot{m}_t^i \cdot T_t^{da}$ term in (7) makes the model non-linear with respect to the control variables. Finally, all zone-specific temperature prediction models (6) are combined to form the multi-zone building model $\mathcal{F}$ in (5). Model parameters a^{ij} and $\mathbf{b}^{ij}$ are derived from system identification (i.e., learned from building historical data).

III. CONTROLLER DESIGN

In this section, we present the design of the proposed RL controller and an MPC baseline. Basics of RL are not discussed here, and interested readers should refer to [16].

A. Reinforcement Learning Controller (RLC)

To be solved by RL, the optimal control problem (5) is first transformed into a Markov decision process (MDP), which is typically represented by a quintuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$. Except the state transition probability $\mathcal{P}$, which is implicitly determined by the controlled system and the environment, other elements $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma)$ are defined as follow.

First, the RL state and its complete set ($\mathbf{s}_t \in \mathcal{S}$) represent the information that the RL agent requires to make a decision. The state typically contains information regarding the current system status and some other information related to its future evolving trajectory. In this study, $\mathbf{s}_t = [\mathbf{T}_t, \mathbf{T}_t^{out}, \mathbf{E}_t, \overline{\mathbf{P}}_t, t, \mathbf{w}_t]$ is used:

1) $\mathbf{T}_t \in \mathbb{R}^N$ are the zone temperatures at t.

2) $\mathbf{T}_t^{out} = [T_{t-K+1}^{out}, ..., T_t^{out}] \in \mathbb{R}^K$ is the T_t^{out} trajectory of the previous K steps. In contrast to some prior works, where a forecast of T_t^{out} is used, we use past data and let the RL agent learn to predict future T_t^{out} implicitly, which avoids developing a forecasting module that adds extra complexity.

3) $\mathbf{E}_t = [f_t, sin_t, cos_t]$ approximates the solar radiation and occupancy level to avoid deploying a sophisticated sensor network within the building. The binary $f_t \in \{0, 1\}$ indicates if the day is a weekday and sin_t, cos_t are the sine and cosine representations for time of the day (e.g., 12:00 PM is represented by $[sin(2\pi \cdot 12/24), cos(2\pi \cdot 12/24)]$). This coarse approximation, inevitably, introduces error; but we will demonstrate it is adequate in a building with a regular schedule in Section V.

4) $\overline{\mathbf{P}}_t = [\overline{P}_t, ..., \overline{P}_{t+K}] \in \mathbb{R}^K$ is the power limit for the next K steps. This means DR events, if called, are notified K steps before their start, which is realistic (e.g., the capacity bidding program from San Diego Gas & Electric "Day-of" option will notify customers hours before the event starts [17]).

5) Finally, the current step t and multi-objective weights $\mathbf{w}_t$ are included in $\mathbf{s}_t$ as well.

Second, the definition of action and its complete set are straight-forward: $\mathbf{a}_t = [\dot{m}_t^1, \dot{m}_t^2, ..., \dot{m}_t^N, T_t^{da}] \in \mathcal{A} \subseteq \mathbb{R}^{N+1}$, corresponding to control variables in (5). Similarly, the immediate reward is defined according to (5) as well: $r_t = -\mathbf{w}_t[\kappa_1 \sum_{i \in \mathcal{N}} \mathcal{D}(T_t^i), \kappa_2 \mathcal{E}_t, \kappa_3 \mathcal{V}_t]^\top \in \mathcal{R}$ to be maximized. Finally, a discount factor $\gamma = 0.99$ is used.

To sum up, based on the MDP formulation above, an RL agent will learn from experience a control strategy (formally introduced in Section IV-A as the *policy*) that maximizes the expected episodic discounted reward: $\mathbb{E}(\sum_{t \in \mathcal{T}} \gamma^t r_t)$. Note that although this MDP problem is different from the optimal control problem (5), it approximates it well as will be shown in Section V.

B. Model Predictive Controllers (MPC)

Two versions of MPC, MPC-ROM and MPC-LIN, are developed to solve (5) based on the ROM developed in Section II-B. The two versions serve as a reference against which the performance of the RLC is compared to. Specifically, MPC-ROM uses the exact ARX model (6), meaning the model used to compute the optimal control actions is the same as the simulation model. In contrast, MPC-LIN is based on a linear approximation of (6), which is obtained by performing a Taylor series expansion at every time-step. Since MPC-ROM requires solving a non-convex optimization, the computational requirements make it less feasible for real-world applications as the number of control variables increases. Therefore, in this paper, MPC-ROM serves as a best possible reference for the control performance; in contrast, MPC-LIN shows the performance of a more attainable implementation.

It is also worth noting that MPC requires exogenous inputs for the planning horizon. In this paper, perfect forecasts (PF) of T_t^{out}, $Q_t^{solar,i}$ and $Q_t^{int,i}$ are provided to MPC-ROM and MPC-LIN, albeit unrealistic. In contrast, RLC only uses historical data of T_t^{out} and coarse approximation (CA) of $Q_t^{solar,i}$ and $Q_t^{int,i}$ for decision making; recall the definition

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

TABLE II

COMPARISON BETWEEN RLC AND MPC

	RL	MPC-LIN	MPC-ROM
Exogenous Data	CA	PF	PF
System Model Knowledge	None	Approximate	Exact

of RL state $\mathbf{s}_t$ in Section III-A. Table II summarizes the comparison between the RLC and MPC settings.

IV. GLOBAL-LOCAL POLICY SEARCH SCHEME

In this section, a global-local policy search scheme is developed, which marries merits from two different types of RL algorithm.

A. RL Algorithm Preliminaries

RL algorithms are generally categorized as value-based or policy-based. Policy gradient methods, as policy-based approaches, directly learn a parameterized control policy ($\pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathbf{s})$, in which $\boldsymbol{\theta}$ is a parameter vector) that maximizes a performance measure $J(\boldsymbol{\theta})$ [16], which is typically represented by $\mathbb{E}_{\pi_{\boldsymbol{\theta}}}(\sum_{t\in\mathcal{T}}\gamma^t r_t)$. At each learning iteration, the RL algorithm will generate a stochastic estimation of the gradient (i.e., $\widehat{\nabla} J(\boldsymbol{\theta})$) from experience and then update $\boldsymbol{\theta}$ using gradient ascent as in:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{8}$$

Two state-of-the-art policy-based model-free RL algorithms, which have proved to have outstanding performance in many benchmark problems, are used in this study. Evolution strategies RL (ES-RL) [12], which uses a zero-order gradient approximation to estimate $\widehat{\nabla} J(\boldsymbol{\theta})$, can achieve fast *global* policy learning due to its gradient-free nature and the ability to escape local optima. Proximal Policy Optimization (PPO) [13] considers KL divergence during learning to avoid destructive large policy updates and thus provides reliable *local* learning.

B. Two-stage Global-Local Policy Searching

Despite the good performance of ES-RL and PPO, using either of them individually might suffer from some drawbacks. Specifically, training an optimal policy using PPO can be slow due to the back-propagation gradient computation, conservative policy update and poor scalability. On the other hand, due to the adoption of the zero-order gradient approximation, ES-RL is likely to converge to the vicinity of the global optimum [18]. In summary, the strength and weakness of these two algorithms are summarized in Table III. Considering that ES-RL and PPO have complementary advantages, we propose a two-stage global-local policy searching scheme to marry the merits from both algorithms. In the first stage, taking advantage of its gradient-free feature and excellent scalability, ES-RL is utilized for fast policy convergence to the vicinity of the global optimum. In the second stage, $\pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathbf{s})$ is transferred to a PPO learner, which locally fine-tunes it using computed gradient information and pushes the policy closer to the global optimum. By utilizing this two-stage framework, the goal is to train an optimal RL controller in a shorter amount of time and with less computational resources when compared with using these algorithms individually. Next, knowledge transfer between two stages is discussed.

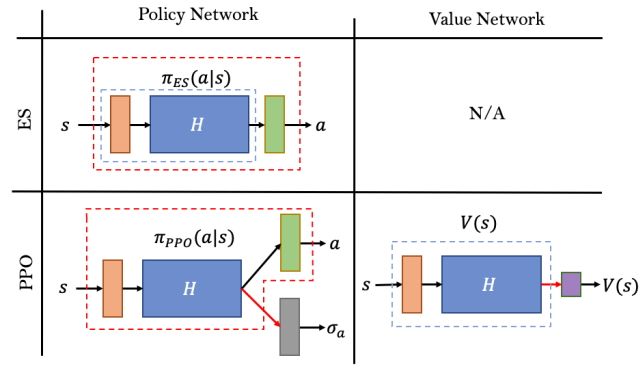


Fig. 1. Weights of the ES policy network are directly copied to the corresponding portion of the PPO policy network (i.e., red dashed box) to warm-start the PPO learner. Similar weight copying is conducted for the PPO value network (i.e., blue dashed box).

C. PPO Learner Warm-Starting

Since ES-RL and PPO both employ neural networks to represent the parameterized policies $\pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathbf{s})$, as long as networks share the same/similar structure, the parameters from a trained ES-RL policy can be directly copied to a PPO network ($\boldsymbol{\theta}_0^{PPO} = \boldsymbol{\theta}^{ES,*}$), which enables the PPO learner to start with nearly as good performance as the well-trained ES learner (though not exactly, as we will explain later). After the weight copying warm start, the PPO learner will further improve $\pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathbf{s})$ using a gradient-based approach. Fig. 1 illustrates the warm-start procedure. It is worth noting that one additional output of the PPO policy network outputs is the standard deviation for each action (denoted as $\sigma_{\mathbf{a}}$). The PPO learner then samples the actual action at each step by a Gaussian distribution $\mathbf{a}_t = \mathcal{N}(\mathbf{a}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix with vector $\sigma_{\mathbf{a}}$ as diagonal elements. Because neural network weights relate to this additional outputs do not exist in the ES network (i.e., the red arrow in the PPO policy network in Fig. 1), we manually initialize the corresponding weights and biases properly to encourage adequate exploration instead of being satisfied with the existing ES policy. Additionally, PPO, as an actor-critic algorithm, uses a value network. We initialize this network using the ES network except for the last layer, due to the different outputs dimension ($V(s)$ is a scalar while $\mathbf{a}$ is a vector). Weights in the last layer are randomly initialized (i.e., red arrow in PPO value network in Fig. 1).

V. CASE STUDY

A. Experiment Setup

A standard EnergyPlus five-zone small office model is used in this paper, see Fig. 2. Parameters of the ROM are obtained by system identification using simulated building operation data. The HVAC power consumption model (2)

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

TABLE III

STRENGTHS AND WEAKNESS OF ES-RL AND PPO

	ES-RL	PPO
Pros	- Scalable - Gradient-free (no back-propagation, fast) - More likely to escape from local optimum	- Policy update with KL divergence considered (stable) - Gradient-based (better local search ability)
Cons	- Likely to converge only to the *vicinity* of global optimum [18]	- Slower learning (back-propagation, conservative update) - Not scalable ($\mathcal{O}(N^2)$ communication complexity of full gradient info) - Gradient-based methods are prone to be trapped in local optimum

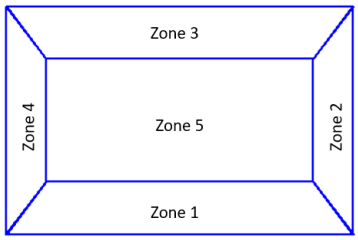


Fig. 2. A five-zone building widely used in building control studies.

has parameters of $a = 1.0$, $b = 0.0076$ and $c = 4.8865$. Control variables ranges are: $T_t^{da} \in [10.0, 16.0]$ ($^\circ C$), and $\dot{m}_t^i \in [0.22, 2.2]$ if $i \neq 5$, otherwise, $\dot{m}_t^5 \in [0.32, 3.2]$ (kg/s). We use monetizing factor $\kappa_i = 1.0$ for all three objectives. Exogenous data from EnergyPlus (Austin, TX), including T_t^{out}, $Q_t^{solar,i}$ and $Q_t^{int,i}$ ($i \in \mathcal{N}$), from July and the first ten days of August are used for training and controller testing respectively. We formulate the building control problem as a episodic MDP with control horizon length of 24 hours and 5-minute control interval ($\mathcal{T} = \{1, 2, ..., 288\}$). The dimension of $\mathbf{T}_t^{out}$ and $\overline{\mathbf{P}}_t$ is 48 (i.e., $K = 48$), which means 1) T_t^{out} for the past 4 hours are used in the state; and 2) a DR event, if called, will be notified 4 hours ahead.

During training, for each episode, we assume the chance a DR event occurs is 50% (50% is higher than reality now, but provided with affordable smart controllers, it is assumed that more frequent DR events will be acceptable in the future). DR events happen anytime between 11:00 and 18:00 with a duration of d minutes and a power limit $\overline{P^{DR}}$ kW. The relationship between d and $\overline{P^{DR}}$ is given by:

$$d = 120 \times (\chi + 1), \quad \overline{P^{DR}} = 30 + 20\chi, \tag{9}$$

where χ follows uniform distribution $U(0, 1)$. In real-life applications, this relationship can be changed appropriately. Given the DR starting time and duration, $\mathcal{T}_{DR} \subseteq \mathcal{T}$ is determined; for episodes without DR events, $\mathcal{T}_{DR} = \emptyset$. In an episode, for $t \notin \mathcal{T}_{DR}$ (normal operation), we use $\mathbf{w}_t = [0.7, 0.2, 0.1]$ and $\overline{P}_t = 80$kW; and when $t \in \mathcal{T}_{DR}$, we use $\mathbf{w}_t = [0.5, 0.0, 0.5]$ and $\overline{P}_t = \overline{P^{DR}}$.

The policy network $\pi_{\boldsymbol{\theta}}(a|s)$ employs the following structure: [108, 256, 128, 128, 64, 64, 32, 16, 6 (ES-RL) / 12 (PPO)] (The first and last items are input and output layers, and the rest show the number of neurons in each hidden layer). Activation functions used in the input and hidden layers are 'tanh' and no activation function is used for the output layer. The RL policy training in this study are conducted on a high-performance computing (HPC) system at the U.S. Department of Energy's National Renewable Energy Laboratory (NREL). Choosing an HPC system as the computing hardware is to help evaluate the training cost for future real-world applications, when controllers are trained on commercial cloud computing platforms.

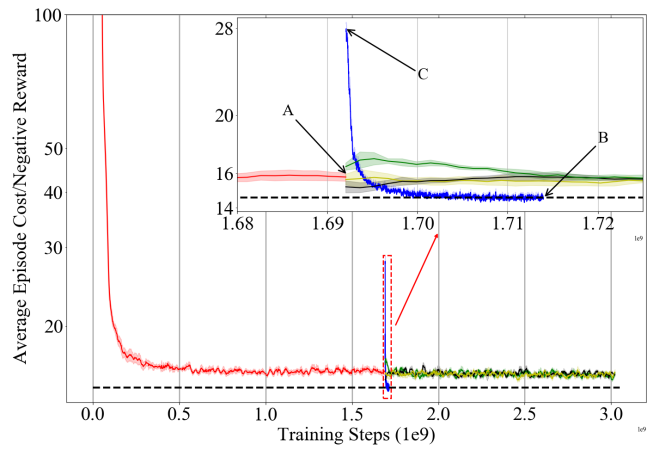


Fig. 3. Learning curves showing the relationship between average episodic cost and training steps: first stage ES curve (red), second stage PPO fine-tuning curve (blue) and three ES fine-tuning curves with different learning step sizes (green: 5×10^{-6}, yellow: 10^{-5} and black: 10^{-6}). All curves are averaged from five individual runs and shaded areas show the standard deviation. Red dashed box highlights the PPO fine-tuning, and is zoomed in and shown at the upper right corner. Black horizontal dashed line shows the converged value after PPO fine-tuning.

B. Two-stage Policy Searching

In the first stage, twenty computing nodes with 683 parallel workers (this number is determined by the available processors on computing nodes) are leveraged for the ES-RL global policy search. In Fig. 3, the red curve shows effective learning since it reveals a cost-reducing general trend as the training progresses. The minimized cost converged within one-hour of run time (in fact, the convergence occurred within half an hour), implying a converged control policy parameterized by $\boldsymbol{\theta}^{ES,*}$. With the trained ES policy network in hand, in the second stage the PPO learner is warm-started as described in Section IV-C for policy fine-tuning, and the training is conducted by 35 parallel workers on one computing node. The learning curve is shown in blue in Fig. 3. It is worth noting that though $\boldsymbol{\theta}^{ES,*}$ is copied exactly, the PPO learner starts the average episodic cost from point 'C' instead of point 'A' in Fig. 3, showing a slightly deteriorated

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

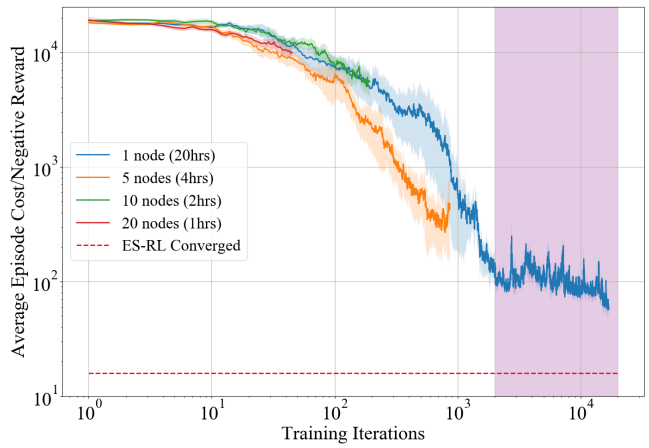


Fig. 4. Learning curves of learning a control policy using PPO without the first stage ES-RL training. Four learning curves showing the PPO learning process using 1 HPC computing node, 5 nodes, 10 nodes and 20 nodes. All experiments are limited with 20 node-hour of computation resource and the learning rate is equal to the one used in ES-RL (i.e., α_1). Again, all curves are averages from five individual runs.

performance at the beginning. Two reasons for this are: 1) the σ_a part of the PPO policy network is initialized to encourage adequate exploration instead of being satisfied with the existing ES policy; thus the PPO learner might explore bad actions as well, which leads to higher cost. 2) the last layer of the PPO value network is randomly initialized, which provides incorrect state evaluation during the first few training iterations. This also leads the PPO learner to have poorer behavior. Despite a start with higher cost, the PPO learner eventually learns a policy with an average episodic cost lower than that achieved by the ES policy, demonstrating an effective policy fine-tuning. When searching the policy globally and locally, learning rates used are $\alpha_1 = 10^{-2}$ and $\alpha_2 = 5 \times 10^{-6}$ respectively.

In this experiment, the two-stage policy search consumes 24 node-hours computational resources (20 nodes for one hour for ES-RL and one node for four hours for PPO) and 5 hours wall time for training the RL controller. Note the first stage convergence happens within the first half hour, it is possible to further reduce the node-hour resources needed.

C. Necessity of the Two-Stage Framework

In this section, considering reasons presented in Section IV-B, we provide empirical evidence that the policy trained by the two-stage learning framework cannot be achieved by using either algorithm alone.

First, gradient-free ES-RL is not effective for fine-tuning. We tested ES-RL with learning rates similar to α_2. However, the three ES fine-tuning curves in Fig. 3 show that even with a reasonable learning rate, using ES-RL for fine-tuning does not yield desirable results. Second, directly using PPO for global policy search can be slow and might be trapped in local optima. Given the same computational resources for the first stage learning in the previous example (i.e., 20 node-hours), Fig. 4 shows the learning curves of using PPO for global policy search with different node numbers. Two observations are made: a) scaling PPO to have more parallel workers is not beneficial: more parallel workers do not lead to more accurate policy gradient computation and faster convergence; in contrast, the increase communication burden (i.e., $\mathcal{O}(N^2)$) might slow the learning down. b) even for the best performing case (i.e., 1 node), PPO converges to a local minimum and is trapped there for the latter 90% of the training time (purple shaded area, note x-axis is in log).

D. RLC Control Performance

To show the effectiveness of the proposed RLC for building control during normal operation and DR events, we simulate two cases where there is/isn't a DR event for a same testing day. Fig. 5 shows the control performance over the 24-hour control horizon for both cases. In the DR case, the building HVAC system is required to limit its power consumption below 36 kW ($\chi = 0.3$) between 14:00 and 16:36. The RLC examined is the PPO fine-tuned one ('B' in Fig. 3).

In Fig. 5, for the non-DR case (dashed lines), $T_t^i(\forall i, \forall t)$ are kept within the comfort band, and P_t is maintained below $\overline{P}_t$, with a peak demand of 56.40 kW. On the other hand, in the DR case, upon receiving the DR notification (four hours prior to the event starts), it can be seen that there are pre-cooling activities: first, T_{da} is reduced to accelerate cooling; second, regarding the control of $\dot{m}_t^i$, we observed a clever learned behavior: the RLC does not mindlessly increase all $\dot{m}_t^i$ since the fan power will increase drastically due to the cubic term in (2); instead, it categorizes zones into three groups and treats them differently:

1) Zone 4: it faces west and receives great solar heat gain during the DR event. Thus, cooling it *during* a DR event is essential and it is not pre-cooled.

2) Zone 2 and 5: east-facing and the core zone are usually not impacted by the afternoon solar heat gain. The RL agent learned from experience that even without pre-cooling, reducing airflow rates in these zones to minimum during DR events will not jeopardize thermal comfort.

3) Zone 1 and 3: they usually require certain amount of cooling in the afternoon. But they will be the focus of pre-cooling in order to shift the cooling load prior to the event.

During the DR event, T_{da} is increased to reduce chiller power consumption and most of the cooling air flow goes to the west-facing Zone 4, while $T_t^i(\forall i, \forall t)$ are mostly within the comfort band and the load reduction requirement is successfully achieved.

E. Comparison with Model Predictive Controllers

The ES-RL trained controller (extracted from 'A' in Fig. 3, denoted as 'RLC-ES'), PPO fine-tuned RLC (extracted from 'B' in Fig. 3, denoted as 'RLC-PPO') and the two MPCs introduced in Section III-B are compared using all ten testing days, and the results are shown in Fig. 6 and Table IV. Both MPCs are implemented using the $fmincon$ solver in MATLAB with the interior-point algorithm.

The following observations are made:

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

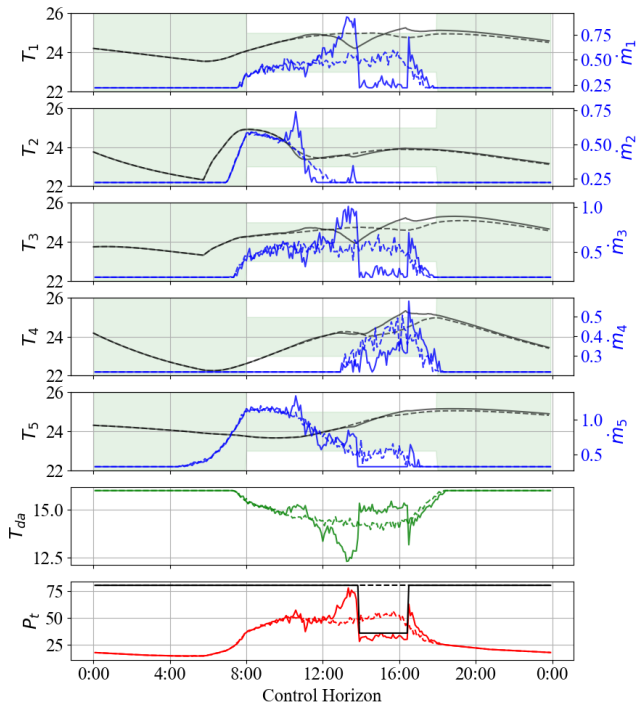


Fig. 5. Top five plots show the indoor temperature profiles (black) and the air flow rate ($\dot{m}^i$) profiles (blue) for five zones. Shaded green areas show the temperature comfort band (The band spans [23, 25] and [22, 28] during occupied and unoccupied period, respectively). The sixth plot shows T_{da} and the last plot shows P_t (red) and $\overline{P}_t$ (black). In all plots, dashed lines represent the non-DR day profiles and solid ones show those of the DR day's.

TABLE IV

COMPARISON OF AVERAGE COST FOR TEST DAYS

Average Cost	RLC-PPO	RLC-ES	MPC-LIN	MPC-ROM
DR	16.31	18.79	18.50	13.75
Non-DR	16.50	17.28	17.70	15.26

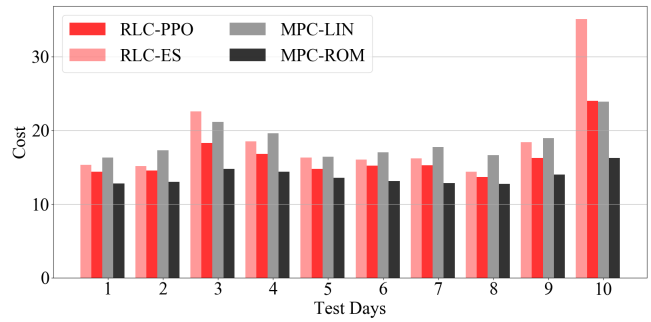
1) Both RLCs can achieve performance similar to or better than that of the MPC-LIN.
2) The second stage PPO fine-tuning is effective as it reduces the RLC-ES's cost by 13.2% and 4.5% for DR scenarios and non-DR scenarios respectively.

It is worth reemphasizing that both MPC-LIN and MPC-ROM use perfect forecasts for exogenous data (i.e., T_t^{out}, $Q_t^{solar,i}$ and $Q_t^{int,i}$) and have either approximate or exact prior knowledge on the building being controlled; in contrast, RLC relies only on easily-accessible historical data for control and has zero knowledge on the building model. Overall, this comparison shows that RLC can be a practical and powerful alternative to MPC in real-world building control applications.

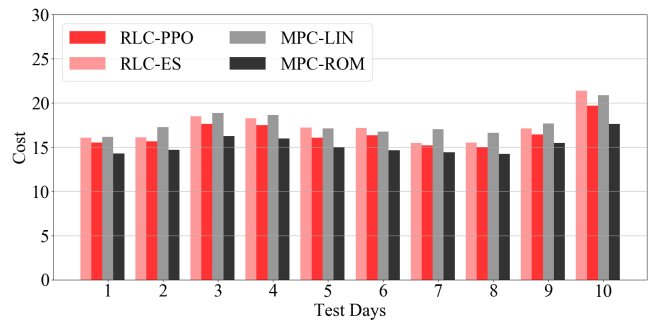
VI. CONCLUSIONS

We demonstrated training an RL controller for a grid-interactive multi-zone office building using the proposed global-local policy search scheme. We showed that the RL controller can optimally control a multi-zone building, under both normal operation and more complicated scenarios such as when providing grid services. Additionally, continuous action space is considered and only easy-to-measure inputs are used for decision making to avoid prohibitive sensor deployment. The experiments reveal that by chaining two different state-of-the-art policy searching paradigms, a control policy can be efficiently trained in the two-stage process. The trained policy, when tested on unseen scenarios, shows proper intelligent control behavior that is entirely learned from the RL agent's trial-and-error experience collection. Moreover, we showed that the RL controller can achieve a similar performance metrics when compared with two different MPC controllers, whose inputs are perfect forecast information.

In this study, different exogenous data are used in controller training and its performance evaluation (i.e., testing using unseen scenarios), however, we assume the building model is accurate (i.e., perfect system identification). In the future, we will consider the discrepancy between the model used in building simulator and the actual building model, and discuss how to train an effective controller using inaccurate building model.



(a) DR Scenarios



(b) Non-DR Scenarios

Fig. 6. Comparing RLCs with MPCs.

REFERENCES

[1] U.S. Energy Information Administration. How much energy is consumed in U.S. residential and commercial buildings? https://www.eia.gov/tools/faqs/faq.php?id=86&t=1. Accessed: Sept. 15th, 2020.

[2] S Somasundaram, RG Pratt, BA Akyol, N Fernandez, N Foster, S Katipamula, E Mayhorn, A Somani, A Steckley, and Z Taylor. Reference guide for a transaction-based building controls framework. *Pacific Northwest National Laboratory*, 2014.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.

[3] U.S. Energy Information Administration. Commercial Buildings Energy Consumption Survey (CBECS) Table E3. Electricity consumption (Btu) by end use, 2012. https://www.eia.gov/consumption/commercial/data/2012/c&e/cfm/e3.php. Accessed: Sept. 15th, 2020.

[4] Yudong Ma, Francesco Borrelli, Brandon Hencey, Brian Coffey, Sorin Bengea, and Philip Haves. Model predictive control for the operation of building cooling systems. *IEEE Transactions on control systems technology*, 20(3):796–803, 2011.

[5] Xiaodong Hou, Yingying Xiao, Jie Cai, Jianghai Hu, and James E Braun. Distributed model predictive control via proximal Jacobian ADMM for building control applications. In *2017 American Control Conference (ACC)*, pages 37–43. IEEE, 2017.

[6] Abdul Afram and Farrokh Janabi-Sharifi. Theory and applications of HVAC control systems–A review of model predictive control (MPC). *Building and Environment*, 72:343–355, 2014.

[7] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[8] Tianshu Wei, Yanzhi Wang, and Qi Zhu. Deep reinforcement learning for building HVAC control. In *Proceedings of the 54th Annual Design Automation Conference 2017*, pages 1–6, 2017.

[9] Zhiang Zhang, Adrian Chong, Yuqi Pan, Chenlu Zhang, and Khee Poh Lam. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy and Buildings*, 199:472–490, 2019.

[10] Naren Srivaths Raman, Adithya M Devraj, Prabir Barooah, and Sean P Meyn. Reinforcement Learning for Control of Building HVAC Systems. In *2020 American Control Conference (ACC)*, pages 2326–2332. IEEE, 2020.

[11] Xiangyu Zhang, Dave Biagioni, Mengmeng Cai, Peter Graf, and Saifur Rahman. An Edge-Cloud Integrated Solution for Buildings Demand Response Using Reinforcement Learning. *IEEE Transactions on Smart Grid*, 2020.

[12] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[14] 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. EnergyPlus: creating a new-generation building energy simulation program. *Energy and buildings*, 33(4):319–331, 2001.

[15] Rohit H Chintala and Bryan P Rasmussen. Automated multi-zone linear parametric black box modeling approach for building HVAC systems. In *Dynamic Systems and Control Conference*. American Society of Mechanical Engineers, 2015.

[16] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[17] San Diego Gas & Electric. Schedule CBP (Capacity Bidding Program). https://www.sdge.com/sites/default/files/Capacity\%20Bidding\%20Program\%20Tariff_0.pdf. Accessed: Sept. 20th, 2020.

[18] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

This report is available at no cost from the National Renewable Energy Laboratory at www.nrel.gov/publications.