energies

MDPI

*Article*

# Reinforcement Learning-Based School Energy Management System

**Yassine Chemingui [1,*], Adel Gastli [1] and Omar Ellabban [2]**

[1] Electrical Engineering Department, Qatar University, P.O. Box 2713 Doha, Qatar; adel.gastli@qu.edu.qa
[2] Iberdrola Innovation Middle East, Qatar Science & Technology Park, P.O. Box 210177 Doha, Qatar; oellabban@iberdrola.com
* Correspondence: yassine.chemingui@qu.edu.qa

Received: 10 September 2020; Accepted: 25 November 2020; Published: 1 December 2020


**Abstract:** Energy efficiency is a key to reduced carbon footprint, savings on energy bills, and sustainability for future generations. For instance, in hot climate countries such as Qatar, buildings are high energy consumers due to air conditioning that resulted from high temperatures and humidity. Optimizing the building energy management system will reduce unnecessary energy consumptions, improve indoor environmental conditions, maximize building occupant's comfort, and limit building greenhouse gas emissions. However, lowering energy consumption cannot be done despite the occupants' comfort. Solutions must take into account these tradeoffs. Conventional Building Energy Management methods suffer from a high dimensional and complex control environment. In recent years, the Deep Reinforcement Learning algorithm, applying neural networks for function approximation, shows promising results in handling such complex problems. In this work, a Deep Reinforcement Learning agent is proposed for controlling and optimizing a school building's energy consumption. It is designed to search for optimal policies to minimize energy consumption, maintain thermal comfort, and reduce indoor contaminant levels in a challenging 21-zone environment. First, the agent is trained with the baseline in a supervised learning framework. After cloning the baseline strategy, the agent learns with proximal policy optimization in an actor-critic framework. The performance is evaluated on a school model simulated environment considering thermal comfort, $CO_2$ levels, and energy consumption. The proposed methodology can achieve a 21% reduction in energy consumption, a 44% better thermal comfort, and healthier $CO_2$ concentrations over a one-year simulation, with reduced training time thanks to the integration of the behavior cloning learning technique.

**Keywords:** energy efficiency; energy management; indoor air quality; reinforcement learning; smart building; thermal comfort

## 1. Introduction

Arid climate prevails in the Arabian Gulf region characterized by mild, pleasant winters; hot, humid summers; and sparse rainfalls. In Qatar, summer temperatures exceed 45 °C and, on average, high temperature exceeds 27 °C in the rest of the seasons. Therefore, air conditioning (AC) in Qatar is more of a necessity than a luxury and accounts for about 80% (highest in the world) of buildings energy consumption. The AC systems are running nonstop throughout the year to maintain thermal comfort. To achieve sustainable development and a greener environment, energy efficiency plays a crucial role in which heating, ventilation, and air conditioning (HVAC) control paves the way forward [1]. However, drastic energy consumption reductions deteriorate the indoor comfort quality, which posits a comfort versus consumption dichotomy. The main goal and challenge of any energy management system (EMS) is to achieve the right balance between occupant comfort and building energy requirements.

Though the occupants' comfort depends on various factors, it is commonly reduced to thermal comfort, for which HVAC controllers are usually optimized, but air quality is seldom considered [2]. In practice, the carbon dioxide ($CO_2$) levels are commonly used as internal air quality (IAQ) indicators. $CO_2$ levels relate to human health and productivity. A school study in [3] concluded that signs of headache and dizziness were prevalent in classrooms with high $CO_2$ levels. Additionally, student performances were better in lower $CO_2$ level environments.

To address these problems, building control algorithms were the subject of extensive research. The methods include classical control, predictive control, and intelligent control [4]. Despite the recent advances in intelligent and predictive controls, the classical on/off and PID control methods [5,6] are the most implemented in the field, due to their simplistic nature. However, these procedures cannot account for the system complexity, stochastic nature, and nonlinear dynamics. Predictive methods try to solve these inherent issues but require complex building modeling and rely on experts' knowledge [7,8]. Therefore, they are hard to generalize over various building environments. Intelligent control methods, instead, are learning-based and model-free, and hence they do not assume complex models underlying the building systems. In these control methods, optimal policies are derived based on collected data, thus alleviating the daunting process of designing a complex mathematically accurate model, which makes them less affected by modeling inaccuracies.

One of such learning-based methodologies is the reinforcement learning (RL), which is a model-free framework for solving optimal control problems stated as Markov Decision Processes (MDPs) [9]. RL is considered the most suitable machine learning paradigm for this task. Building control matches with RL, since there are an environment to control, hidden dynamics to learn, and serial decisions to determine. RL has gained a lot of attention in the past few years due to successes in playing Atari games and then beating the world Go champion. The combination of neural networks as function approximators and RL paradigm was the key. Since then, RL was considered as a viable solution for diverse control problems, in particular, building energy management systems. Previous attempts were limited to tabular RL and RL algorithms using linear function approximators. RL discipline is not new and its applications in building control are not either. Previously, RL algorithms were constrained to computationally cheap algorithms such as tabular Q-learning and linear function approximators and were forced to consider small state/action spaces.

This paper aims to deliver an optimum solution to a multi-objective and multizone building energy management (BEM) problem that provides a comfortable indoor environment in a school building while reducing its energy consumption and, hence, lowering its operational costs. This study leverages a deep reinforcement learning (DRL) framework to develop an artificially intelligent agent capable of handling the tradeoffs between building indoor comfort and energy consumption. To the best of the authors' knowledge, this study is the first to apply a DRL-based, behavioral-cloning-enhanced technique to resolve the interactions between thermal comfort, energy consumption, and indoor air quality in a multi-zone complex environment (21 zones). The experiments were conducted in a school environment in Qatar. The proposed solution handles the tradeoff between energy consumption and occupants' comfort well. The proposed intelligent control can generalize over different weather conditions throughout the year while maintaining good thermal comfort, excellent indoor air quality, and saving more than 20% of the school's energy consumption, compared to a rule-based baseline control strategy.

The main contributions of this work are summarized as follows:

- Propose a proximal policy optimization (PPO) algorithm for energy optimization and occupants' comfort control for maintaining occupant's comfort while reducing energy consumption.
- Use behavioral cloning to incentivize the basic baseline behavior so that the proposed algorithm converges faster than trying very random decisions.
- Develop a complex 21-zone school simulation system with EnergyPlus and thoroughly investigate the performances through meticulously designed experiments.

This paper is organized as follows. Apart from the introduction (Section 1), a literature review is presented in Section 2. Section 3 describes the RL approach and its application to the school case study environment. Section 4 presents and discusses the obtained results. The concluding remarks are given in Section 5.

## 2. Literature Review

The particularity of the Arabian Gulf region climate led to several studies on improving energy efficiency in a desert climate. Buildings account for the majority of energy consumption due mostly to the cooling needs. Building retrofitting was proposed to reduce old building energy demand [10], optimal control of AC was investigated [11], and a multi-objective genetic algorithm was investigated in a Qatari house setup [12]. Until recently, energy efficiency was not considered in the region. With the fall of oil prices, the local governments started raising awareness and designing efficiency programs. The arid environment is indeed challenging, but the highly subsidized electricity tariffs and the limited financial incentives hinder the efforts. Authors in [13] analyzed electricity load profiles in Qatar. They found that approximately 50% of energy demand is attributed to cooling in summer. In [14], the impact of retrofitting and behavioral changes on energy consumption in Abu Dhabi was discussed. It is crucial to raise awareness among the community, since most present buildings do not conform with the efficiency guidelines, and citizens use cooling 24/7 even when the building is empty. Thus, there is a crucial need for strategies to decrease buildings' electricity consumption while maintaining their residents' comfort.

The recent breakthroughs of RL [15–17], due to the powerful combination of deep learning and RL algorithms in game playing, got the attention of and spurred multiple research interests [18,19]. Before, tabular Q-learning and variants were widely applied as RL-based controllers for energy optimization [20–27]. For instance, [20] Q-learning was employed to lower energy consumption by 10% compared to programmable control. Authors in [25] coupled an autoencoder with Q-learning to reach less consumption by 4–9% in winter and 9–11% in summer in contrast to constant set-point policy. Authors in [26] applied State–Action–Reward–State–Action (SARSA) to control the environment based on fixed setpoints and reduce energy consumption, while [27] relied on linear approximation for state—action value. These methods are incapable of ingesting large state/action space. The DRL union handles the dimensionality curse better, replacing tabular search and simple function approximators with neural networks. DRL process high dimensional raw data without the need for preprocessing and feature engineering based on raw data, and hence can accomplish end-to-end control [28–34].

In [28], the authors applied both tabular and batch Q-learning with a neural network to realize a 10% lower energy consumption compared to rule-based control. In [31], a mixture of Long Short Term Memory (LSTM) neural network and actor–critic architecture achieved around 15% thermal comfort improvement and a 2.5% energy efficiency improvement when compared to fixed strategies. Predicted mean vote (PMV) was used as the thermal comfort indicator, and the testbed was one zone office space with two days of simulation for training and five for validation. Authors in [30] compared DQN, regular Q-learning, and rule-based on/off control in reducing the HVAC consumption and maintaining a prespecified comfortable temperature (24 °C). The algorithm was evaluated on three simulated buildings with EnergyPlus [35] (one-zone, four-zone, and five-zone models). The DQN bested the other methods with over 20% energy reduction. Authors in [36] resorted to the Asynchronous Advantage Actor–Critic (A3C) algorithm, which was trained to reduce energy and ensure good thermal comfort, measured by the predicted percentage of dissatisfied (PPD) index, in a simulation of a workplace building in Pittsburgh. Fifteen percentage of energy consumption was reduced compared to their base case. Similar to our work in [33], the authors used double Q-learning to optimize for IAQ and thermal comfort while reducing energy consumption by 4–5% in a laboratory and classroom simulation setup. To the best of the authors' knowledge, and based on the literature review, zones considered in the reviewed papers and studies do not exceed five zones [30]. These papers also focused on thermal comfort and usually defined it as fixed temperature preferences.

In this work, the building's energy consumption, thermal comfort, and indoor air quality in a more complex environment comprising of 21 zones have been optimized, where the agent selects the optimal decisions from $72^{21}$ possible action combinations at each time step.

Previous attempts applied DQN [15] variants and its continuous extension deep deterministic policy gradient (DDPG) [37]. However, PPO algorithms [38], the leading policy search algorithms, have not been studied in the context of energy efficiency. PPO has shown promising results in physical control problems providing more stable learning and simpler hyperparameter tuning than previous policy gradient algorithms. PPO achieved close or above state of the art performance on a wide range of tasks, becoming the default RL algorithm at openAI. In this paper, we apply PPO to control a school building simulation and achieve excellent indoor comfort and significantly reduce energy consumption.

## 3. Proposed RL Methodology

In contrast to the known machine learning paradigms, RL deals with sequential decision making under uncertainty. In supervised learning, the data is labeled, and thus the right decisions are previously known, whereas in RL setup, the artificial agent learns from experience. Based on scalar feedback, it updates its behavior through trial and error. Different from unsupervised learning, the agent has the reward feedback. Furthermore, the RL agent generates data and experience while understanding the environment: in this work, the simulated school via EnergyPlus. The goal in RL is to maximize future returns. The agent searches for the optimal sequence of decisions. When judging a situation, the agent takes into account the possible future effects of the current decision.

To develop the right strategy, the agent explores the environment depending on these essential components:

- States describing agent and/or environment position.
- Actions affecting the environment.
- Rewards as feedback from the environment on the chosen action.

Figure 1 describes the interactive process between the agent and the environment.

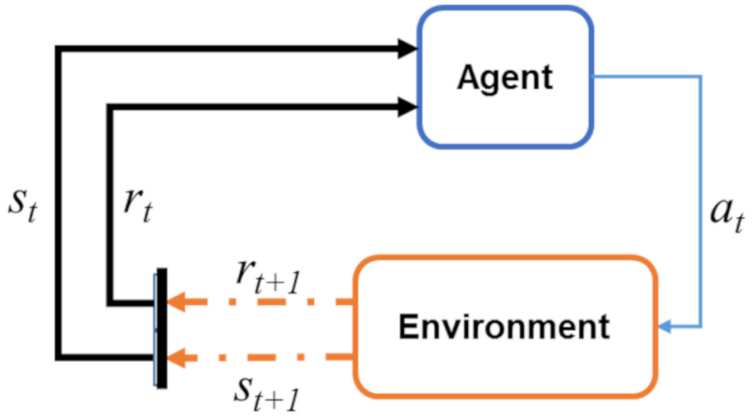


**Figure 1.** Reinforcement Learning Framework.

MDPs are the mathematical framework for RL. An MDP is a tuple of states s, actions a, reward function r, transitioning probabilities p, and discount factor $\gamma$: $<s, a, p, r, \gamma>$. Usually, due to the lack of knowledge of environment dynamics ($p$), model-free methods to estimate the value and policy functions are used. Value functions assess how good the current state ($V(s)$) or state/action couple ($Q(s,a)$) is. The policy can be derived by selecting the actions that maximize the Q value (act greedily). Alternatively, via policy gradient algorithms, the policy can be optimized directly. Value-based methods learn the optimal policy by deriving the value function like in Q-learning. In contrast, policy-based methods estimate the optimal strategy directly, like in REINFORCE [39]. The policy parameters are optimized. Actor–critic is a combination of both methods.

As illustrated by Figure 2, in actor–critic methods, both the policy and value functions are estimated in order to learn a good policy. The actor represents the policy, while the critic represents the value function. The critic estimates guide the learning with the temporal difference (TD) error $(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$. The general update rule is given by Equation (1):

$$V(s_t) = V(s_t) + \alpha \left(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right) \quad (1)$$

where $V$ is the value function, $s_t$ is the state at time $t$, $\alpha$ is the stepsize, $\gamma$ is the discount factor, and $r_t$ is the reward at time $t$. If the error is positive, the current behavior is encouraged, and the probability of selecting the recent action increases by means of the policy gradient theorem.

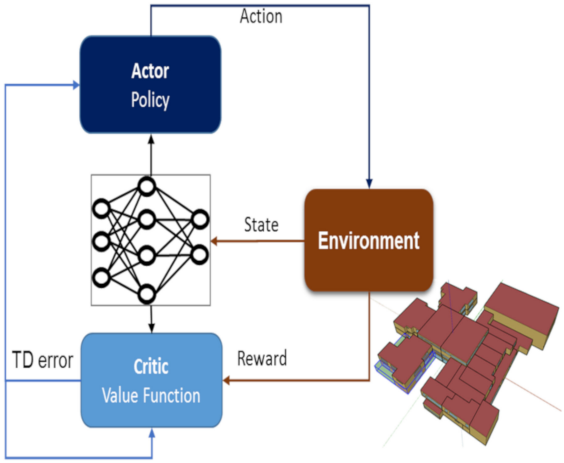


**Figure 2.** Actor–critic architecture.

DDPG achieved good results in continuous control tasks. However, selecting the right hyperparameters is tricky. This is common in policy gradient methods. Trust region policy optimization (TRPO) algorithms iteratively optimize policies while guaranteeing improvement over the old policy [40]. TRPO algorithms are on-policy algorithms, where the agent's behavior is updated according to its current behavior. They are more stable than DDPG, and they relax the difficulty of choosing a precise step size with fewer hyperparameters tuning. Constrained to a certain degree of improvement from the old policy to the new one, the policy is updated modestly with small changes at a time via maximizing a surrogate objective, as shown in Equation (2):

$$\max_{\theta} \mathbb{E}_{\pi_{old}}\left[\frac{\pi(a_t|s_t)}{\pi_{old}(a_t|s_t)} A^{\pi_{old}}(s_t|a_t)\right] \text{subject to } \mathbb{E}\left[KL[\pi(.|s_t),\ \pi_{old}(.|s_t)]\right] < \delta \quad (2)$$

where $\pi$ is the policy (actor) function, which is the probability of selecting $a_t$ given $s_t$, A is the advantage, it helps reduce variance $A(s,\ a) = Q(s,\ a) - V(s)$, and KL is Kullback–Leibler divergence. Policy changes are constrained by $\delta$, and the difference between old and new policies is measured in terms of Kullback–Leibler divergence.

TRPO has its disadvantages too. The monotonic improvement costs heavy computations to calculate the Fisher Matrix and conjugate gradient from KL divergence. In the same year, TRPO's leading author also proposed proximal policy optimization algorithms to alleviate computation and conserve TRPO's stability. Since PPO has become open AI's default algorithm. Despite its simplicity, it achieves performance comparable, and sometimes even better than state-of-the-art approaches. The most interesting feature of PPO is the ease of tuning, a characteristic rarely seen in RL research. In PPO, the surrogate objective is clipped. The policies ratio $r_t$ is constrained to the range of $[1 - \varepsilon; 1 + \varepsilon]$ to limit fluctuations between old and new strategies ($\varepsilon$ is a hyperparameter).

The objective function selects then the lower bound or the pessimistic estimate, as shown in Equation (3) and Equation (4).

$$L^{CLIP}(\theta) = \mathbb{E}[\min(r_t(\theta)A_t, clip(r_t(\theta),\ 1-\varepsilon,\ 1+\varepsilon)A_t)] \tag{3}$$

$$r_t = \frac{\pi(a_t|s_t)}{\pi_{old}(a_t|s_t)} \tag{4}$$

Table 1 shows the algorithm of the PPO using actor–citric style.

**Table 1.** PPO pseudo-code.

| **PPO Algorithm** |
|---|
| 1: Initialize policy parameters $\theta_0$ and value function parameters $]\varphi_0$ |
| 2: for k = 0,1,2, . . . . Do |
| 3: Collect set of trajectories $D_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment |
| 4: Compute rewards-to-go $R_t$ and advantages estimates ${A_t}^{\pi_{\theta_k}}$ |
| 5: Update the policy: $\theta_{\mathbf{k}+1} = \underset{\theta}{\mathbf{argmax}} L^{CLIP}_{\theta_k}(\theta)$ |
| 6: Update the value function: $]\boldsymbol{\varphi}_{\mathbf{k}+1} = \underset{]\varphi}{\mathbf{argmin}} \mathbb{E}\left[V_{]\varphi}(s_t) - R_t\right]^2$ |
| 7: end for |

### *3.1. Behavior Cloning*

Behavior cloning (BC) is a form of imitation learning in which the agent learns a policy through supervised learning. The proposed algorithm collects an expert's knowledge or behavior, usually a combination of state–action pairs. The data is then fed to the agent to force the expert's behavior. This a supervised learning task. The agent is trained to match states with actions. In the proposed methodology, demonstrations are gathered from the simulation following the baseline strategy. Similar to RL, a decision is made, and the environment reaction is documented. The contrast here is that decisions are based on the baseline behavior, not drawn from the agent's policy. The same interaction pipeline is run, and the filed data is stored. The resulting state–action pairs are used for training the agent in order to mimic the initial cloned strategy. This is to force the agent to follow the baseline decisions and use them later on as a benchmark. Hence, when selecting a new action, it evaluates its potential compared to the baseline. We gather information from a year of simulations, then the baseline behavior is cloned before the training of the agent in the usual trial and error framework of RL. The advantage of BC is that the agent learns the desired behavior without interacting with the environment. Subsequently, the agent interacts with the environment as predefined and searches for better policies. The difference is, instead of starting with random unreliable actions in the exploration phase, the agent has the baseline behavior to build upon it as ground truth. Thus, erratic behaviors are avoided and training time is diminished.

### *3.2. School Testbed Control Framework*

A simulated environment was developed based on a real school in Qatar, which is considered the case study testbed. The school architecture, a typical Qatari school, was organized into 21 zones, which were selected based on their common air conditioning configuration and control. The zones correspond to classrooms, offices, laboratories, and other facilities. The school layout is presented in Appendix A.1, specifying the zones. For instance, the air handling unit (AHU) 17 controls the gym, AHUs 8 to 15 control classrooms, and AHU 4 controls the hall. The simulation embeds the school's orientation and exposure to the sun and also the weather of the region. The RL agent is trained and evaluated using this testbed with typical weather conditions covering a whole year. The simulation sampling time is 15 min. EnergyPlus is used for this task.

EnergyPlus is a fully integrated building and HVAC simulation program developed by the U.S. Department of Energy. It models buildings, heating, cooling, lighting, ventilating, and other energy

flows. It is used also for load calculations from energy use, modeling natural ventilation, photovoltaic systems, thermal comfort, water use, etc. Besides energy consumption, the simulation software tools can also be used to calculate the following variables:

- Indoor temperatures
- Needs for heating and cooling
- Consumption needs of HVAC systems
- Natural lighting needs of the occupants
- Interior comfort of the inhabitants
- Levels of ventilation

As shown in Figure 3, the first step is to construct the 3D modeling of the building with the SketchUp software.

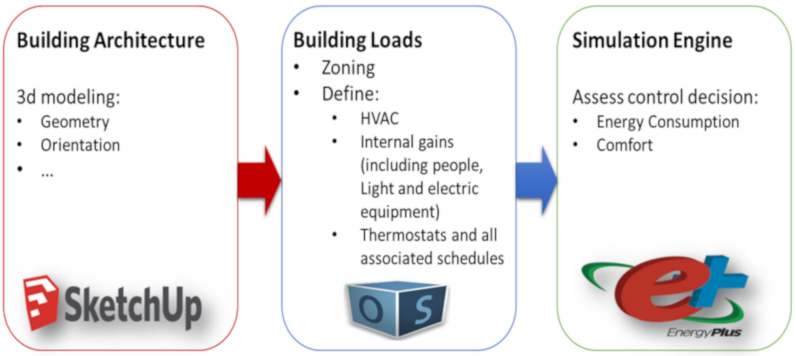


**Figure 3.** Simulation design.

Then the various zones are defined with their loads and their controls with the OpenStudio software. Finally, the model is exported to an ".idf" file, which is the file format used by the EnergyPlus software as a building model under study.

Once the modeling is finished, a Python program is developed for co-simulation. The proposed framework is developed in Python, and the communication between the EnergyPlus and the agent is provided by the PyEp library [41]. The intelligent controller is composed of two multilayer perceptron (MLP) networks; one for the actor and the other for the critic. The neural networks are developed in PyTorch. Each network is simply comprised of only two hidden layers of size 256 with ReLU activations Equation (5).

$$\mathrm{ReLU}(x) = \max(0, x) \tag{5}$$

Adam optimizer [42] is applied with a learning rate of 3e-4. As shown in Figure 4, at every time step, based on the environment state, the agent estimates the state-value function (critic) on one hand. On the other hand, it decides the optimal course of action (actor). Then, it receives a feedback signal and adjusts its behavior accordingly.

### *3.3. Baseline*

During working hours, the temperature is set to 21 °C and 28 °C when the building is unoccupied. The $CO_2$ levels are maintained under 1000 ppm at night and under 700 ppm during the day.

### *3.4. States*

At every timestep, the agent observes the environment to construct the state and act upon it. The state comprises the temperature, relative humidity of each zone, the outside temperature,

and relative humidity, and the time step information. We opted for minimal information to ensure the ease of implementation in the real world. The state $s_t$, at time $t$, is then determined using (6).

$$s_t = (t, T_{outside}, H_{outside}\ T_{zone\ 1},\ H_{zone\ 1}, \ldots\ldots,\ T_{zone\ 21},\ H_{zone\ 21}) \quad (6)$$

All these variables are normalized to the range of [0,1].

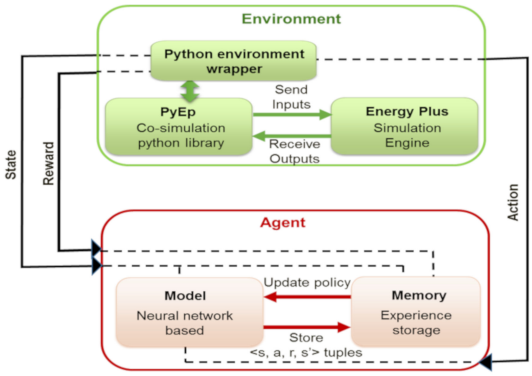


**Figure 4.** Environment and agent interactions process.

### 3.5. Actions

The actor $a_t$ decides for each zone the setpoints of the temperature (°C) and $CO_2$ (ppm), as shown in (7).

$$a_t = (Tset_{zone\ 1}, CO_2set_{zone\ 1},\ \ldots\ldots, Tset_{zone\ 21}, CO_2set_{zone\ 21}) \quad (7)$$

Note that there are $(12 \times 6)^{21}$ action combinations:

- 12 temperature setpoints, from 17 °C to 28 °C
- Six $CO_2$ setpoints: from 500 ppm to 1000 ppm.

### 3.6. Reward

The reward at any time $t$ is a scalar value, $r_t$ designed in a way to motivate the optimal behavior. The objective is to reduce energy consumption and maintain good thermal comfort and indoor air quality. Therefore, the reward is composed of two terms: energy-related and comfort-related terms, as in (8) and (9), where $\alpha$ and $\beta$ are both taken equal to 0.5.

$$r_t = -\alpha\ .\ energy_t - \beta\ .\ discomfort_t \quad (8)$$

$$discomfort_t = thermal\ discomfort_t + hygienic\ discomfort_t \quad (9)$$

### 3.7. Comfort

Comfort is divided into two categories: thermal and hygienic comforts.

#### 3.7.1. Thermal Comfort

Comfort is defined here by means of the predicted mean vote (PMV). PMV is an index, developed by Fanger, that aims to predict the mean value of votes of a group of occupants on a seven-point thermal sensation scale, as shown in Figure 5.


**Figure 5.** Predicted mean vote (PMV).

PMV is based on heat-balance equations and empirical studies about skin temperature to define comfort. Thermal equilibrium is obtained when an occupant's internal heat production is the same as its heat loss. PMV equal to zero is representing thermal neutrality. Fanger's equations are used to calculate the PMV of a group of subjects for a particular combination of air temperature, mean radiant temperature, relative humidity, airspeed, metabolic rate, and clothing insulation. PMV is a rigorous index for comparing the performances of different approaches. Since the PMV is a robust measure and its values are easily understandable, we chose it taking into account the model deployment and tuning later on. In a real-world implementation, it will be replaced by the occupant's feedback. The occupant will select a value from the PMV seven points. We hypothesize that the PMV reflects well enough occupant's comfort. Since the reward is a scalar feedback signal, we reduce the comfort to the average over the zones. Lower values suggest good comfort, and thus we evaluate discomfort as the absolute value of the average. The thermal comfort interval of [-0.5,0.5] is considered optimal; therefore, no penalties are incurred by the agent.

In the present study, the thermal discomfort is calculated using Equation (10):

$$discomfort = \begin{cases} 0, & |PMVavg| < 0.5 \\ |PMVavg|, & |PMVavg| \geq 0.5 \end{cases} \quad (10)$$

where

$$PMVavg = \frac{1}{21} \sum_{zone=1}^{21} PMV_{zone} \quad (11)$$

Figure 6 illustrates the relationship between the discomfort and the PMV average value as defined by Equation (10) and Equation (11).

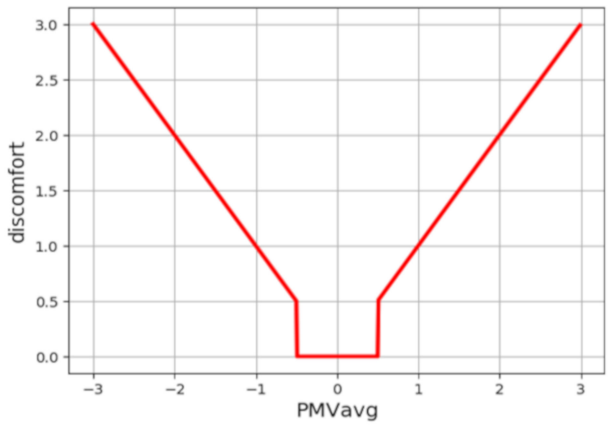


**Figure 6.** Thermal discomfort.

### 3.7.2. Hygienic Comfort

The hygienic comfort or discomfort is measured in terms of the indoor $CO_2$ levels using Equation (12).

$$discomfort = \begin{cases} 0, & CO_2avg < 600\ ppm \\ \left[\frac{CO_2avg - 600}{1600}\right]^2, & 600 \leq CO_2avg \leq 1000\ ppm \\ 4, & CO_2avg > 1000\ ppm \end{cases} \quad (12)$$

The optimal $CO_2$ concentrations (good: healthy Levels) are usually within the range of [400 ppm, 600 ppm]. For this range, no discomfort is recorded. Above it, the $CO_2$ concentrations become mediocre and even bad for health (see Table 2). For the [600 ppm, 1000 ppm], we opted for a quadratic discomfort that increases faster than a linear one to emphasize the danger of escalating levels of $CO_2$ concentrations. When levels surpass 1000 ppm, the situation becomes dangerous for human health,

and thus we raise the discomfort dramatically to restrain the agent from reaching those conditions. The hygienic discomfort versus $CO_2$ concentration, according to Equation (12), is illustrated in Figure 7.

**Table 2.** $CO_2$ levels effects on health.

| $CO_2$ (ppm) | Air Quality |
|---|---|
| 400 | Good: Healthy Levels |
| 600 | |
| 800 | |
| 1000 | Mediocre: Drowsiness and Odors |
| 1200 | |
| 1400 | |
| 1600 | Bad: Risk for Health Damage |
| 1800 | |

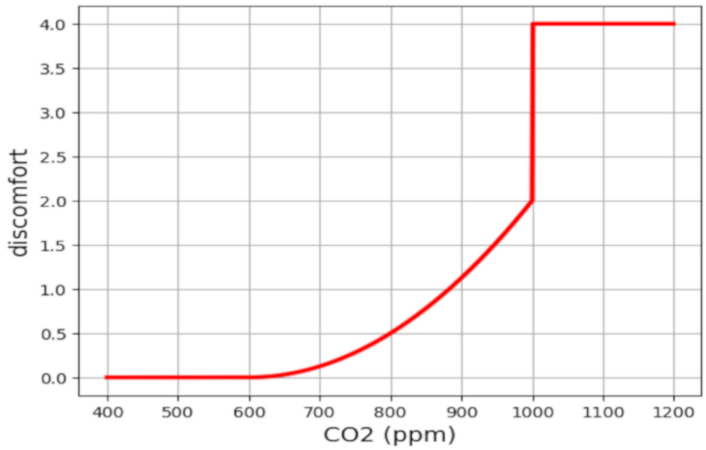


**Figure 7.** Hygienic discomfort.

## 4. Results

During the simulation, each zone has its characteristics, which increases the complexity of the optimization task. As shown in Figure 8, the 21 zones differ in volume. They also differ in their exposure to direct sun radiations and in the sun-facing angle. Therefore, the optimum temperature setting changes are expected to take place in some zones more significantly than in others.

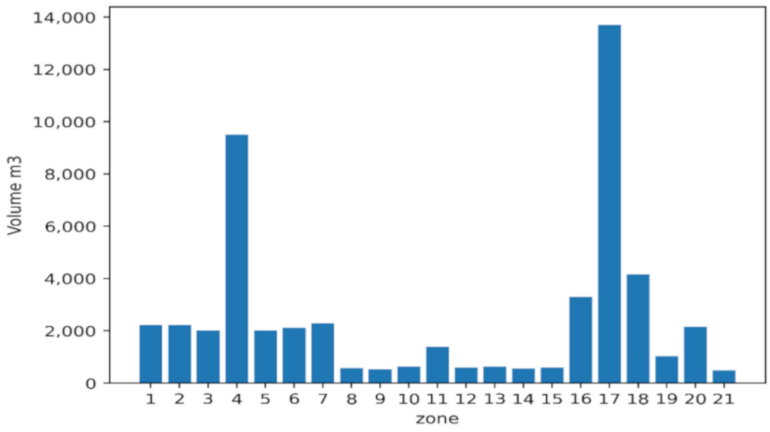


**Figure 8.** Zones volume.

The agent must navigate these variations to find optimal solutions. The dissimilarity is mostly noticed in energy consumption, because the comfort component has the same value ranges across zones. However, reaching the same comfort level for two zones requires different energy levels. Our agent is

trained throughout the year, and its performance is evaluated against the baseline. We stop training when cumulative returns stabilize. An episode is a year of simulation. The agent training was done on an intel core i7-5600U cpu and each episode took around ≈ 10 min. Our results are very promising, taking into account the simplicity of the neural networks shallow architectures (only two hidden layers). Additionally, only temperature and humidity variables were needed as state information. At first, the baseline behavior is cloned to reduce computation time and achieve better results than training with zero knowledge, as shown in Figure 9.

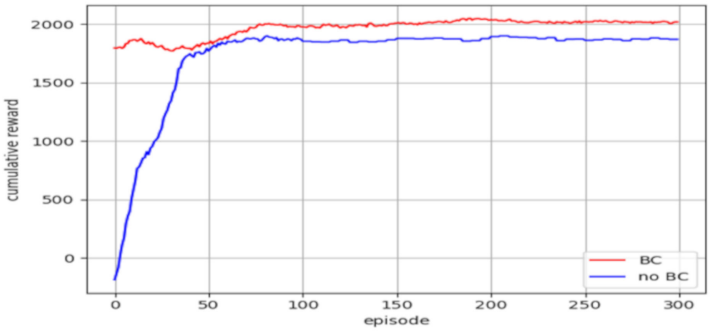


**Figure 9.** Comparison between training with behavior cloning and without.

With BC, we start with more rewards at the beginning of learning and achieve better results in the long term. This is due to the exploration/exploitation tradeoff. The raw agent tries many variants of decisions until it reaches a good strategy. However, with BC, it learns to perform better than a good baseline from the start.

Energy consumption and thermal comfort improvements in different weather conditions are investigated. Energy consumption reduction varies from month to month, but the agent is always capable of decreasing energy consumption, as illustrated in Figure 10.

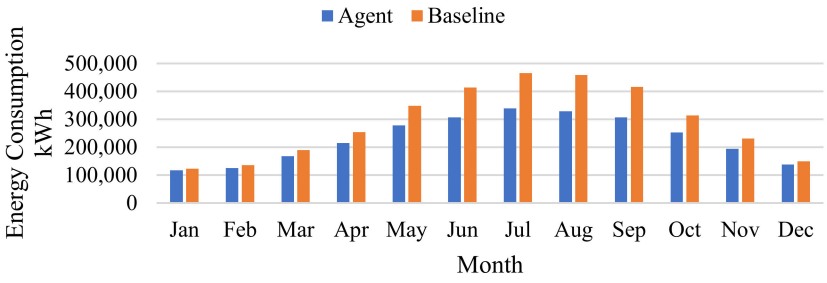


**Figure 10.** Energy consumption comparison between agent and baseline.

For PMV comfort, in some cases, the proposed agent strategy allows less comfort compared to the baseline. This allows for having less energy consumption, while the PMV levels remain mostly inside the [−0.5,0.5] range. Overall, the proposed methodology can achieve a 21% reduction in energy consumption and 44% better thermal comfort. Figure 11 summarizes the monthly gains in comfort and energy consumption for the whole year. The optimized strategy results are compared with those of the baseline in terms of energy and comfort and report the percentage of improvement. It is clear that the amelioration follows the outdoor temperature profile well.

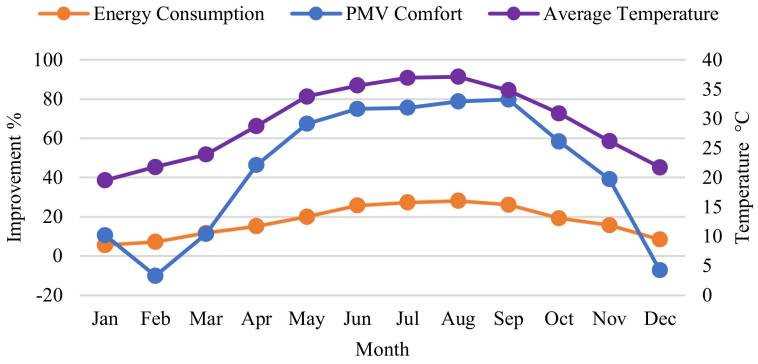


**Figure 11.** Energy and thermal comfort improvement per month.

In August, for instance, a 28% reduction in energy consumption is achieved. During the cold months, energy reduction is less significant, since the outdoor weather is pleasant. The lowest record is obtained in January, during which the energy consumption is optimized by only 6%. Also during the cold months, less improvement in thermal comfort is recorded. For example, in February, the thermal comfort is worse by 10% in terms of PMV. This might sound poor, but the PMV this month is still within the desired range of [−0.5,0.5]. The values outside the range correspond to points in the working day start or end, and they do not stretch over significant periods. Notice that the mean PMV is inside the admissible range for all the zones. Notice also that in the baseline, a cold sensation is present in some zones, even though the overall PMV values are better than the optimization. Arguably, the agent has better comfort, because the baseline reaches bad comfort values, as displayed in Figure 12.

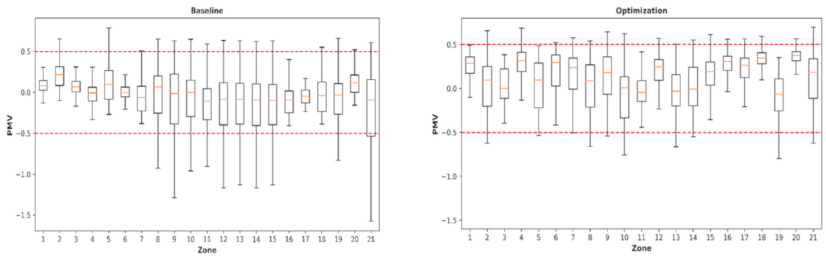


**Figure 12.** PMV in February following: (**a**) baseline's strategy; (**b**) agent's strategy.

Though indoor air quality differs from season to season and from zone to zone, good $CO_2$ levels are consistently maintained in our experiment. $CO_2$ concentrations are always under 1000 ppm, and they are the highest in July, because the agent automatically prioritizes the energy consumption and thermal comfort. During this period, maintaining good thermal comfort with reduced energy is challenging due to the high temperatures. The plots of $CO_2$ levels per zone for four months, a month per season, can be found in Appendix A.2.

In Appendix A.3, the PMV values are presented per zones for four months [a month per season]. The thermal comfort is maintained in the desired range. PMV varies from zone to zone due to their different characteristics. It also varies from season to season. The tendency to the warmer environment due to hot weather is noticeable. Values are within the desired [−0.5,0.5] interval overall. Boxplots depict the data quartiles, and some values outside the optimal range are present. These values are common and do not reach uncomfortable levels. Their span is brief to accommodate for the start or end of the day and optimize energy consumption. The values also follow the weather. For instance, in January, these PMV values are lower overall than in the other months, since in Qatar, the weather is hot throughout the year, and temperatures drop only during the winter.

The comfort variables for the 21 zones and the four seasons (summer, autumn, winter, and spring) are summarized in Figures 13–16 (one figure for one variable: PMV, zone temperature, $CO_2$, and relative

humidity). Notice that, for all the 21 zones and all the seasons, the mean $CO_2$ levels do not exceed 800 ppm. The maximum value is 1000 ppm reached in summer. Therefore, the contaminant concentration is always in the healthy range. The temperatures in summer are obviously lower than in the other seasons. The mean temperature changes from one season to another and does not exceed 2 °C, because the weather does not vary drastically over the year. Though the comfort levels are limited within the [−0.5,0.5] range, the mean values per zone are successfully maintained in the range of [−0.35,0.35]. PMV values rarely exceed the desired range, and when it happens, they remain below the slightly uncomfortable thermal comfort values (+1 or −1). These values correspond to the start and end of the working day, and they do not harm the overall comfort, since the values span short periods.

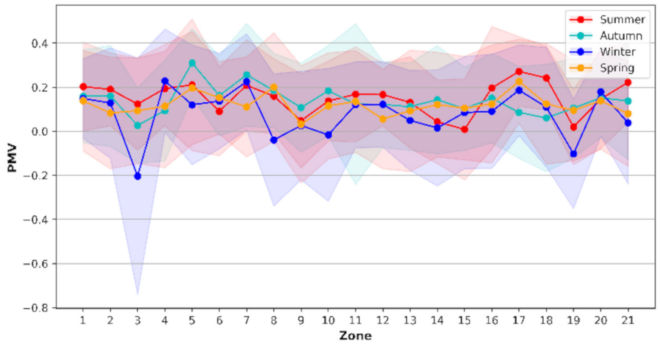


**Figure 13.** PMV mean over the seasons (hues account for standard deviation).

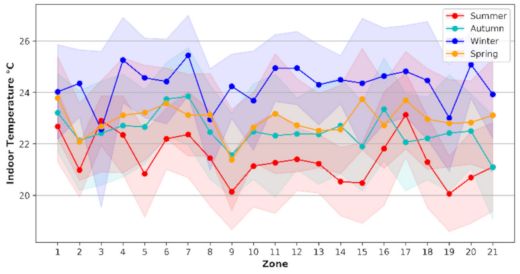


**Figure 14.** Indoor temperatures mean over the seasons (hues account for standard deviation).

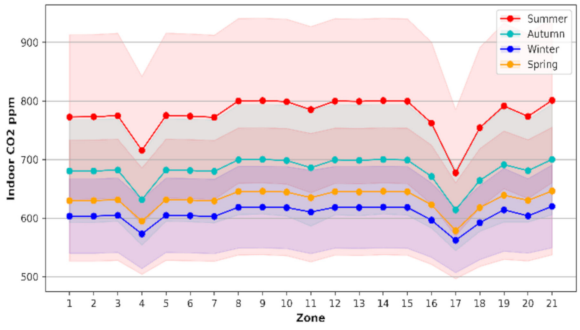


**Figure 15.** $CO_2$ concentrations mean per season (hues account for standard deviation).

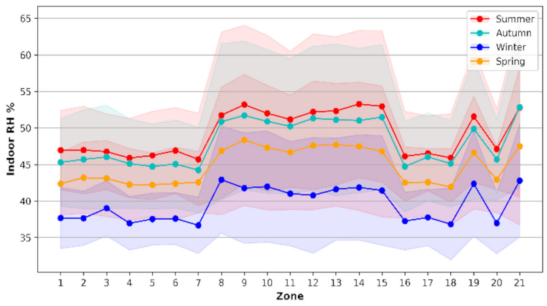


**Figure 16.** Relative humidity mean per season (hues account for standard deviation).

## 5. Conclusions

In this work, deep reinforcement learning is applied to control a school building's indoor environmental conditions. The school building is a 21-zone environment that is modeled and simulated using EnergyPlus. The proximal policy optimization is used to train the intelligent agent. The learning process is sped up by cloning the baseline strategy at the first step before learning new policies. None of the previous studies of DRL control for building energy management applied PPO or behavioral cloning. Additionally, compared to other works, the proposed testbed is the most complex with $72^{21}$ possible actions at every timestep. The agent successfully learns the optimal control decisions for different weather conditions throughout the year. The performance is then evaluated over one year of simulation, achieving a 21% reduction in energy consumption while preserving a very good indoor comfort. More interestingly, the agent achieves such results with shallow neural networks as function approximators.

In the next step, the focus will be on deploying the agent into a real school environment and investigating its performance. In addition, the behavioral cloning effect on learning should be studied in more detail and the transferability of the learned strategy to other environments should be evaluated.

**Author Contributions:** Conceptualization, Y.C.; Formal analysis, Y.C.; Funding acquisition, O.E. and A.G.; Investigation, Y.C.; Methodology, Y.C.; Project administration, A.G.; Software, Y.C.; Supervision, O.E. and A.G.; Validation, Y.C.; Visualization, Y.C.; Writing – original draft, Y.C.; Writing – review & editing, O.E. and A.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Qatar National Research Fund grant number [NPRP10-1203-160008].

**Acknowledgments:** This publication was made possible by the National Priority Research Program (NPRP) grant [NPRP10-1203-160008] from the Qatar National Research Fund (a member of Qatar Foundation) and the co-funding by IBERDROLA QSTP LLC. The findings achieved herein are solely the responsibility of the authors.

**Conflicts of Interest:** The authors declare that there is no conflict of interest.

## Nomenclature

**GLOSSARY OF TERMS**

| **ACRONYMS** | **DEFINITION** |
|---|---|
| A3C | Asynchronous Advantage Actor-Critic |
| AHU | Air Handling Unit |
| BC | Behavior Cloning |
| BEM | Building Energy Management |
| DDPG | Deep Deterministic Policy Gradient |
| DQN | Deep Q Network |
| DRL | Deep Reinforcement Learning |
| EMS | Energy Management System |

| | |
|---|---|
| HVAC | Heat, Ventilation and Air Conditioning |
| IAQ | Internal Air Quality |
| KL | Kullback-Leibler |
| LSTM | Long Short Term Memory |
| MDP | Markov Decision Process |
| MLP | Multilayer Perceptron |
| PID | Proportional, Integral and Derivative |
| PMV | Predicted Mean Vote |
| PPD | Predicted Percentage Dissatisfied |
| PPO | Proximal Policy Optimization |
| RL | Reinforcement Learning |
| SARSA | State–Action–Reward–State–Action |
| TD | Temporal Difference |
| TRPO | Trust Region Policy Optimization |

## Appendix A

### *Appendix A.1 School Architecture*

Figures here show the 3D model that is developed for the case-study school building, the ground floor, and first-floor layouts, respectively.

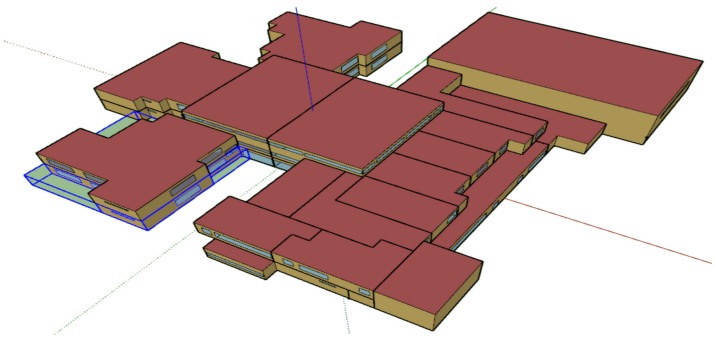

**Figure A1.** School 3D model.

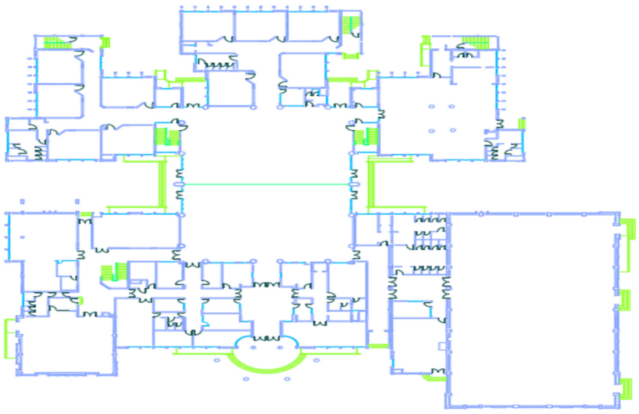

**Figure A2.** School ground floor.

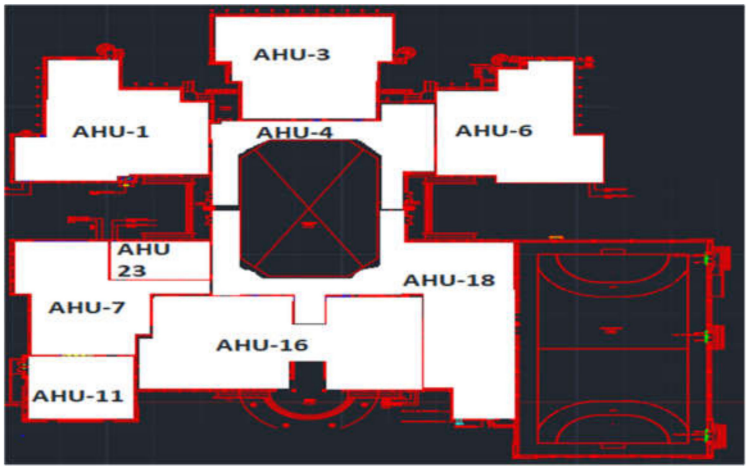


**Figure A3.** Air handling units in ground floor.

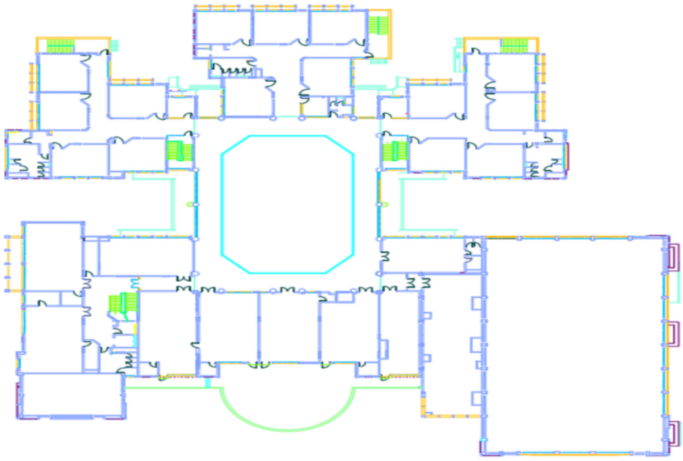

**Figure A4.** School first floor.

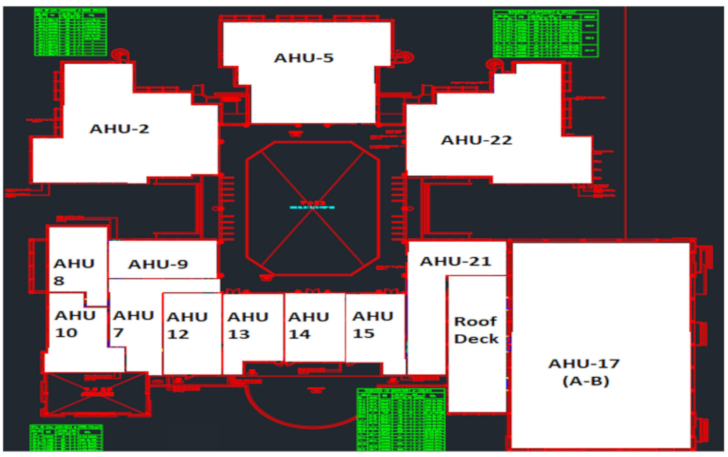


**Figure A5.** Air handling units in first floor.

## *Appendix A.2 $CO_2$ Concentrations per Zone for Four Months*

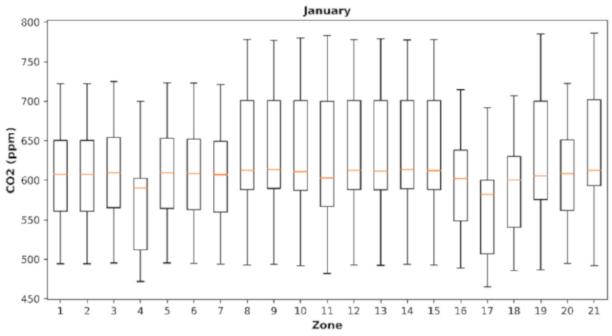


**Figure A6.** $CO_2$ concentrations in January.

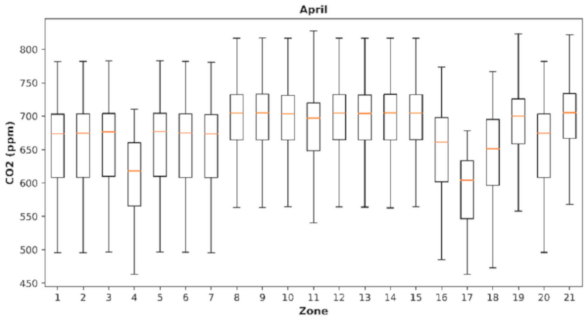


**Figure A7.** $CO_2$ concentrations in April.

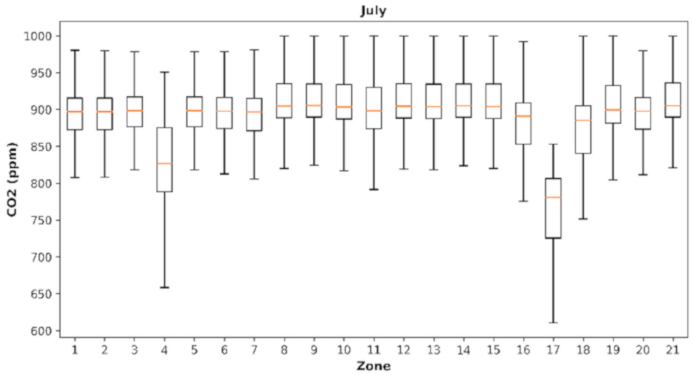


**Figure A8.** $CO_2$ concentrations in July.

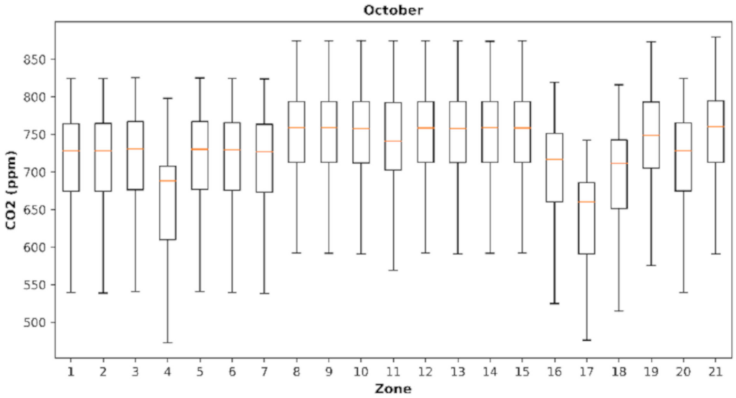


**Figure A9.** $CO_2$ concentrations in October.

*Appendix A.3 PMV Values per Zone for 4 Months*

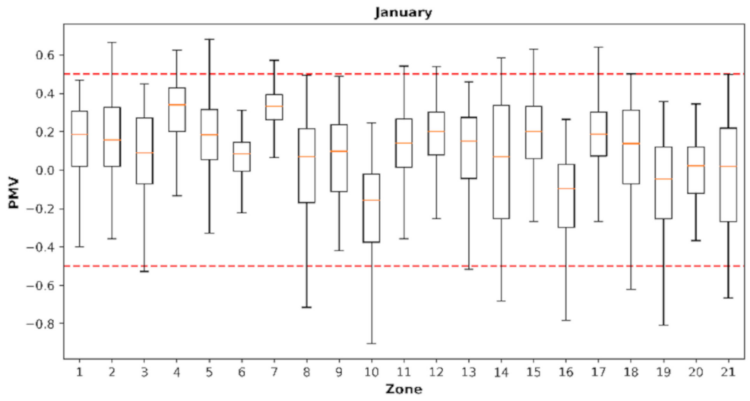


**Figure A10.** PMV in January.

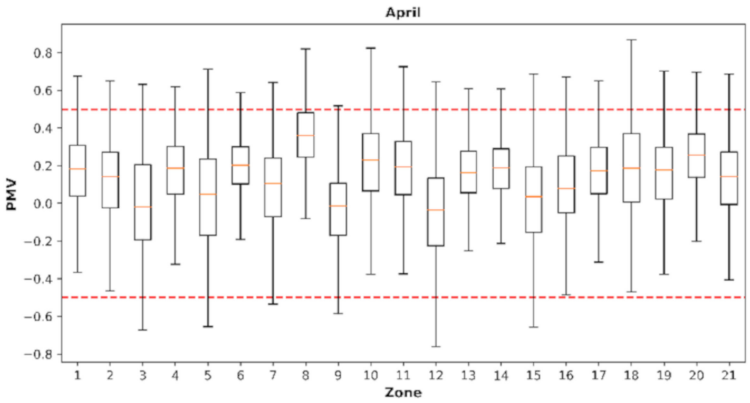


**Figure A11.** PMV in April.

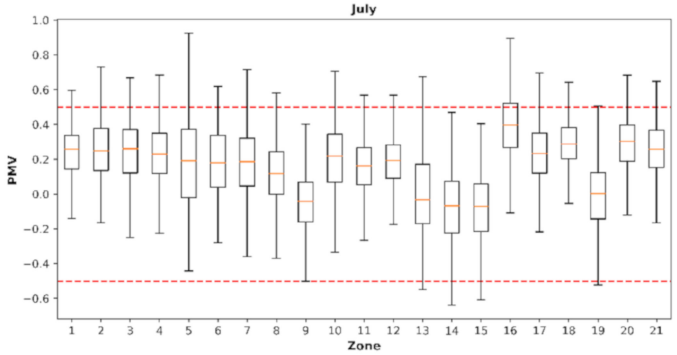


**Figure A12.** PMV in July.

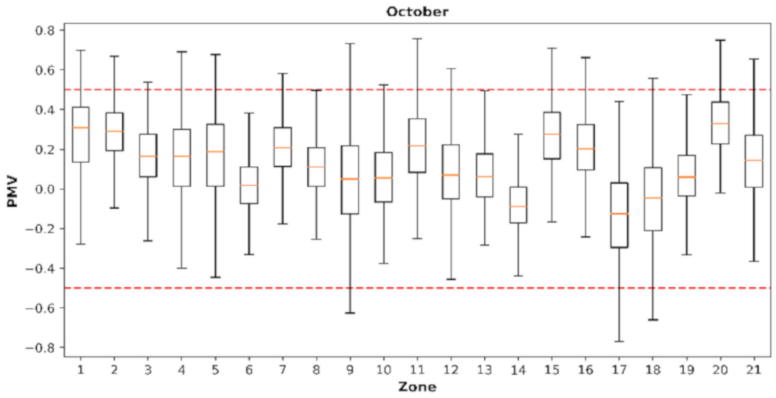


**Figure A13.** PMV in October.

## References


1. Hijawi, U.; Gastli, A.; Hamila, R.; Ellabban, O.; Unal, D. Qatar green schools initiative: Energy management system with cost-efficient and lightweight networked IoT. In Proceedings of the 2020 IEEE International Conference on Informatics and Enabling Technologies (ICIoT), Doha, Qatar, 2–5 February 2020; pp. 415–421. [CrossRef]
2. Han, M.; May, R.; Zhang, X.; Wang, X.; Pan, S.; Yan, D.; Jin, Y.; Xu, L. A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustain. Cities Soc.* **2019**, *51*, 101748. [CrossRef]
3. Myhrvold, A.N.; Olsen, E.; Lauridsen, O. Indoor environment in schools–pupils health and performance in regard to $CO_2$ concentrations. In Proceedings of the 7th International Conference on Indoor Air Quality and Climate, Nagoya, Japan, 21–26 July 1996.
4. Belic, F.; Hocenski, Z.; Sliskovic, D. HVAC control methods—A review. In Proceedings of the 2015 19th International Conference on System Theory, Control and Computing, ICSTCC 2015—Joint Conference SINTES 19, SACCS 15, SIMSIS 19, Cheile Gradistei, Romania, 14–16 October 2015. [CrossRef]
5. Levermore, G. *Building Energy Management Systems*; Routledge: London, UK, 2013.
6. Dounis, A.I.; Bruant, M.; Santamouris, M.; Guarracino, G.; Michel, P. Comparison of Conventional and Fuzzy Control of Indoor Air Quality in Buildings. *J. Intell. Fuzzy Syst.* **1996**, *4*, 131–140. [CrossRef]
7. Ma, Y.; Borrelli, F.; Hencey, B.; Coffey, B.; Bengea, S.; Haves, P. Model Predictive Control for the Operation of Building Cooling Systems. *IEEE Trans. Control. Syst. Technol.* **2012**, *20*, 796–803. [CrossRef]
8. Wei, T.; Zhu, Q.; Maasoumy, M. Co-scheduling of HVAC control, EV charging and battery usage for building energy efficiency. In Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), San Jose, CA, USA, 2–6 November 2014; pp. 191–196. [CrossRef]
9. Puterman, M.L. *Markov Decision Processes*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1994.
10. Felimban, A.; Prieto, A.; Knaack, U.; Klein, T. Energy retrofitting application research to achieve energy efficiency in hot-arid climates in residential buildings: A case study of Saudi Arabia. *World Acad. Sci. Eng. Technol. J. Archit. Environ. Eng.* **2020**, *14*, 185–188.
11. Al-Azba, M.; Cen, Z.; Remond, Y.; Ahzi, S. An optimal air-conditioner on-off control scheme under extremely hot weather conditions. *Energies* **2020**, *13*, 1021. [CrossRef]
12. Benhmed, K.; Ellabban, O.; Gastli, A. Novel home energy optimization technique based on multi-zone and multi-objective approach. In Proceedings of the 2nd International Conference on Smart Grid and Renewable Energy, SGRE 2019—Proceedings, Doha, Qatar, 19–21 November 2019. [CrossRef]
13. Bayram, I.S.; Saffouri, F.; Koç, M. Generation, analysis, and applications of high resolution electricity load profiles in Qatar. *J. Clean. Prod.* **2018**, *183*, 527–543. [CrossRef]
14. Giusti, L.; Almoosawi, M. Impact of building characteristics and occupants' behaviour on the electricity consumption of households in Abu Dhabi (UAE). *Energy Build.* **2017**, *151*. [CrossRef]
15. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]

16. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Driessche, G.V.D.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
17. Berner, C.; Brockman, G.; Chan, B.; Cheung, V.; Dębiak, P.; Dennison, C.; Farhi, D.; Fischer, Q.; Hashme, S.; Hesse, C.; et al. Dota 2 with Large Scale Deep Reinforcement Learning. Available online: http://arxiv.org/abs/1912.06680 (accessed on 19 May 2020).
18. Li, Y. *Reinforcement Learning Applications*; 2019. pp. 1–41. Available online: http://arxiv.org/abs/1908.06973 (accessed on 19 May 2020).
19. Wang, Z.; Hong, T. Reinforcement learning for building controls: The opportunities and challenges. *Appl. Energy* **2020**, *269*, 115036. [CrossRef]
20. Barrett, E.; Linder, S. Autonomous hvac control, a reinforcement learning approach. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science*; Bifet, A., Ed.; Springer: Cham, Switzerland, 2015; Volume 9286. [CrossRef]
21. Li, B.; Xia, L. A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings. In Proceedings of the 2015 IEEE International Conference on Automation Science and Engineering (CASE), Gothenburg, Sweden, 24–28 August 2015; pp. 444–449. [CrossRef]
22. Nikovski, D.; Xu, J.; Monaka, M. A Method for Computing Optimal Set-Point Schedule for HVAC Systems. In Proceedings of the Clima 2013, 11th REHVA World Congress and 8th International Conference on Indoor Air Quality, Ventilation and Energy Conservation in Buildings, Prague, Czech Republic, 16–19 June 2013.
23. Liu, S.; Henze, G.P. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory. *Energy Build.* **2006**, *38*, 142–147. [CrossRef]
24. Henze, G.P.; Schoenmann, J. Evaluation of Reinforcement Learning Control for Thermal Energy Storage Systems. *HVAC R Res.* **2003**, *9*, 259–275. [CrossRef]
25. Ruelens, F.; Iacovella, S.; Claessens, B.; Belmans, R. Learning Agent for a Heat-Pump Thermostat with a Set-Back Strategy Using Model-Free Reinforcement Learning. *Energies* **2015**, *8*, 8300–8318. [CrossRef]
26. Zenger, A.; Schmidt, J.; Krödel, M. Towards the intelligent home: Using reinforcement-learning for optimal heating control. In *KI 2013: Advances in Artificial Intelligence*; Timm, I.J., Thimm, M., Eds.; KI 2013. Lecture Notes in Computer Science, vol 8077; Springer: Berlin/Heidelberg, Germany, 2013. [CrossRef]
27. Dalamagkidis, K.; Kolokotsa, D.; Kalaitzakis, K.; Stavrakakis, G. Reinforcement learning for energy conservation and comfort in buildings. *Build. Environ.* **2007**, *42*, 2686–2698. [CrossRef]
28. Yang, L.; Nagy, Z.K.; Goffin, P.; Schlueter, A. Reinforcement learning for optimal control of low exergy buildings. *Appl. Energy* **2015**, *156*, 577–586. [CrossRef]
29. Vázquez-Canteli, J.; Kämpf, J.; Nagy, Z. Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration. *Energy Procedia* **2017**, *122*, 415–420. [CrossRef]
30. Wei, T.; Wang, Y.; Zhu, Q. Deep reinforcement learning for building HVAC control. In Proceedings of the 54th Annual Design Automation Conference, Austin, TX, USA, 18–22 June 2017; pp. 1–6. [CrossRef]
31. Wang, Y.; Velswamy, K.; Huang, B. A Long-Short Term Memory Recurrent Neural Network Based Reinforcement Learning Controller for Office Heating Ventilation and Air Conditioning Systems. *Processes* **2017**, *5*, 46. [CrossRef]
32. Nagy, A.; Kazmi, H.; Cheaib, F.; Driesen, J. Deep Reinforcement Learning for Optimal Control of Space Heating. Available online: http://arxiv.org/abs/1805.03777 (accessed on 19 January 2020).
33. Valladares, W.; Galindo, M.; Gutiérrez, J.; Wu, W.-C.; Liao, K.-K.; Liao, J.-C.; Lu, K.-C.; Wang, C.-C. Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. *Build. Environ.* **2019**, *155*, 105–117. [CrossRef]
34. Li, Y.; Wen, Y.; Tao, D.; Guan, K. Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning. *IEEE Trans. Cybern.* **2020**, *50*, 2002–2013. [CrossRef]
35. Crawley, D.B.; Lawrie, L.K.; Winkelmann, F.C.; Buhl, W.; Huang, Y.; Pedersen, C.O.; Strand, R.K.; Liesen, R.J.; Fisher, D.E.; Witte, M.J.; et al. EnergyPlus: Creating a new-generation building energy simulation program. *Energy Build.* **2001**, *33*, 319–331. [CrossRef]
36. Zhang, Z.; Chong, A.; Pan, Y.; Zhang, C.; Lam, K.P. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy Build.* **2019**, *199*, 472–490. [CrossRef]
37. Xiang, J.; Li, Q.; Dong, X.; Ren, Z. Continuous Control with Deep Reinforcement Learning. Available online: http://arxiv.org/abs/1509.02971 (accessed on 1 February 2020).

38. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. Available online: http://arxiv.org/abs/1707.06347 (accessed on 1 February 2020).
39. Williams, R.J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* **1992**, *8*, 229–256. [CrossRef]
40. Schulman, J.; Levine, S.; Moritz, P.; Jordan, M.I.; Abbeel, P. Trust Region Policy Optimization. Available online: http://arxiv.org/abs/1502.05477 (accessed on 1 February 2020).
41. PyEp. Available online: https://github.com/mlab-upenn/pyEp (accessed on 19 January 2020).
42. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference Learn Represent ICLR 2015—Conference Track Proceeding, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.