
Zheming Liang††, Student Member, IEEE, Desong Bian†*, Member, IEEE, Chunlei Xu§, Xiaohu Zhang†, Member, IEEE, Di Shi†, Senior Member, IEEE, and Zhiwei Wang†, Senior Member, IEEE

Abstract—More than 30% of the total energy generated all over the world are consumed by buildings. Among it, 40% are consumed by commercial buildings which are great candidates to perform energy management strategy. Nonetheless, comfort levels of the occupants in the commercial building always have higher priority comparing to the energy-consuming cost, especially during the business hours. As a result, coordinating the operation of large appliances and other major components in commercial buildings needs a novel energy management strategy, which provides a trade-off between the comfort level and building operation cost. In this paper, we develop a deep reinforcement learning-based control strategy to determine optimal actions for major components in a commercial building to minimize operation costs while maximizing comprehensive comfort levels of occupants. An unsupervised deep Q-network method is introduced to handle the energy management problem by evaluating the influence of operation costs on comfort levels considering the environment factors at each time slot. An optimum control decision can be derived that targets both immediate and long-term goals, where exploration and exploitation are considered simultaneously. Extensive simulation results show that the proposed deep reinforcement learning-based energy management strategy is capable of both minimizing operation and maintenance costs and maximizing comprehensive comfort levels of occupants.

I. INTRODUCTION

In U.S., buildings consume approximately 40% of the total power supply, where commercial buildings (CBs) account for more than 50% among the total building power consumption [1]. Besides, residential households are too large in amount and too small in size compared with CBs which are not suitable to perform MW level demand response. Moreover, the peak hours of residential households (7:00 am-10:00 am and 5:00 pm-8:00 pm) are completely different from that of the CBs (11:00 am-4:00 pm), where the office hours of CBs share the same time interval as the peak hours of the distribution system. Furthermore, when curtailting or shifting the same amount of power loads, the influenced areas for the residential loads are much larger than the commercial campuses. Thus, we focus on implementing energy management strategy for CBs in this paper.

Nonetheless, comfort levels of the occupants in the CB always have higher priority comparing to the energy-consuming cost, especially during the business hours. As a result, a novel energy management strategy which provides a trade-off between the comfort level and building operation cost is needed to coordinate the operation of large appliances and other major components in CBs. Currently, the energy management strategy can be classified into three different types: (i) to minimize operating costs of the distribution system/microgrids; (ii) to maximize the comfort level of the consumers in the distribution system/microgrids or to minimize the discomfort level of the consumers in the distribution system/microgrids (mainly for comfort/discomfort level related to the indoor temperature); and (iii) to minimize the load curtailment cost/time/influence areas. Only few of prior works have jointly considered all these aspects [2].

Even though the well-known optimization approaches such as stochastic programming can provide the global optimal energy management decisions directly with accurate models, these detailed models are hard to get or are not accurate or changing fast, especially in the distribution network. A dynamic and self-adapting algorithm is necessary to fill such gaps. The current trend of energy management in the distribution system is based on reinforcement learning (RL) [3]. An occupant centered controller for lighting in CBs based on RL is proposed in [4]. However, a discrete state space is not enough to represent the complex energy usage condition for a CB. Moreover, the system model changes every season, which cannot be handled by a set of determined parameters. Thus, the deep reinforcement learning (DRL) algorithm is developed to involve continuous state/action space in the real-world real-time energy management problem. A policy gradient based actor-critic reinforcement learning algorithm is introduced in [5] to provide optimal energy management decisions for CB central controllers. Moreover, a heuristic DRL-based algorithm is developed to maintain desired indoor temperature for distributed buildings in [6]. Again, only indoor temperature related comfort level is considered in the environment.

In this paper, we develop a DRL control strategy to provide...
optimal control decisions for major components in a CB to minimize operation and maintenance costs for the central controller while maximizing occupants’ comprehensive comfort levels. An unsupervised DRL algorithm is developed to handle the energy management problem through evaluating the influence of operation and maintenance costs on comfort levels related environments at each time slot. A deep Q-network method is utilized to provide an optimum control decision through the learning process. The trained neural network can target all aforementioned objectives, where exploration and exploitation are considered simultaneously. Extensive simulation results based on real-world data sets indicate that our proposed framework is able to minimize operating costs of a CB while maximizing consumers’ comprehensive comfort levels.

In summary, the main contributions of this paper are as follows.

1) A deep reinforcement learning based energy management strategy is developed to minimize operation and maintenance costs for the central controller of the CB and maximize comprehensive comfort levels for the occupants simultaneously.
2) A deep Q-network is utilized to provide optimal energy management decisions for the central controller of the CB.
3) Extensive simulation results based on the real-world data sets show the proposed energy management framework is reliable to handle all related uncertainties considering comprehensive comfort levels.

The rest of this paper is organized as follows. In Section II, the models of a CB are described. Section III presents the mathematical formulation of our proposed problem. We conduct a case study in Section IV and then draw conclusions in Section V.

II. SYSTEM MODELING

In the proposed commercial system, there are one CB and one parking lot, which can be scaled to multiple CBs smoothly. In the CB, there are one heating, ventilation, and air-conditioning (HVAC) system, one electric water heater (EWH), one energy storage system (ESS), one solar panel, and one aggregated base power load. There are several plug-in electric vehicles (PEVs) that owned by the occupants in the CB desire to be charged in the parking lot during the office hours. The CB also suffers from uncertain demand response request from the upstream grid operator. In order to ensure the power balance of the commercial system in the whole operating day, several reliability constraints are proposed. There are 96 time slots in each operating day, i.e., each time slot has 15 minutes.

A. Appliances Constraints

The detailed formulation of aforementioned major appliances in the commercial system are described in the following content.

1) HVAC: The HVAC system is one of the most important appliances in a commercial system, especially for CBs that have critical loads that are sensitive to temperature deviations, such as servers. In order to measure the satisfaction level of the critical loads, the indoor temperature has been selected as one of the unique features to represent the comfort level related to the HVAC system. Moreover, the relationship between the dynamic of the power consumption of the HVAC system and the indoor temperature is the key point to address the trade-off between electricity cost and comfort level in the objective function. Thus, we first model the indoor temperature dynamics of the CB as follows:

$$T_{t+1}^{hvac} = \beta^{hvac} T_{t}^{hvac} + \alpha^{hvac} U_{t}, \forall t,$$

where $T_{t}^{hvac} = [T_{t}^{in}, T_{t}^{ow}, T_{t}^{ewh}]^{T}$, including indoor temperature, inner wall temperature, and outer wall temperature, respectively. $U_{t} = [T_{t}^{out}, \Psi_{t}, \sigma_{t}, \rho_{t}^{hvac}]^{T}$, including outdoor temperature, solar irradiance, binary on/off action indicator of the HVAC system, and constant power consumption of the HVAC system, respectively. $\alpha$ and $\beta$ are environment coefficients of CBs [7].

In addition, to ensure that the critical loads cannot be damaged by the indoor temperature, we set the upper and lower bounds of the indoor temperature within the deviation from the desired indoor temperature.

$$T_{d}^{in} - \delta^{hvac} \leq T_{t}^{in} \leq T_{d}^{in} + \delta^{hvac}, \forall t,$$

where $T_{d}^{in}$ is the maximum temperature deviation from the desired indoor temperature.

2) EWH: Despite the requirement to maintain the indoor temperature, the hot water demand within a CB is another unique feature that needs to be tackled. In our model, the hot water demand is satisfied by an EWH, where the hot water can be stored in the water tank attached to the EWH. The water temperature within the hot water tank is selected as the representative for the comfort level related to the EWH. The detailed dynamic relationship between the water temperature and the power consumption are modeled as follows:

$$T_{\tau}^{ewh} = T_{0}^{ewh} + \Delta T_{\tau}^{ewh}, \Delta T_{\tau}^{ewh} = \sum_{t=1}^{\tau} \zeta_{t}^{ewh} \zeta_{t+1}^{ewh} \frac{P_{t}^{ewh} - H_{t}^{de}}{C_{water} M}, \forall \tau,$$

where $P_{t}^{ewh}$ is constant power consumption of the EWH. Binary variable $\zeta_{t}^{ewh}$ denotes the on/off action indicator of the EWH. $\zeta_{ewh}$ is the power-to-heat ratio of the EWH. Auxiliary state variable $\Delta T_{\tau}^{ewh}$ is the temperature deviation of the EWH between the beginning of the operating day and the time $\tau$. Parameter $H_{t}^{de}$ represents the aggregated negative impacts on the temperature of the hot water in the EWH, including heat loss that is transferred to its ambient, outflow of hot water and inflow of cold water. Parameter M is the mass of water in the hot water tank, and $C_{water}$ is the specific heat capacity of water.

3) ESS: Since roof-top solar panels are considered in the proposed model, the uncertainty within the power generation process of the roof-top solar panels need to be mitigated.
Additionally, reliable power supply for the critical loads such as servers must be guaranteed. Thus, the ESS is implemented in the CB to mitigate potential power imbalances. Unlike dynamic energy levels of ESS modeled in conventional studies, we use the state-of-charge (SoC) to represent the energy dynamics within the ESS as follows:

\[ \text{SoC}_{k,t} = \text{SoC}_{k,t-1} + \frac{p_{ch,k,t} v_{ch,k,t} - p_{dis,k,t} v_{dis,k,t}}{E_k}, \forall k, t, \]  

(4)

where \( p_{ch,k,t} \) and \( p_{dis,k,t} \) are constant power charged into or discharged from the \( k \)-th ESS at time \( t \), and \( v_{ch,k,t} \) and \( v_{dis,k,t} \) represent charging and discharging efficiencies of the \( k \)-th ESS, respectively. \( u_{ch,k,t} \) and \( u_{dis,k,t} \) are binary variables indicating charging and discharging decisions of the \( k \)-th ESS. Each ESS has a finite capacity, therefore, energy stored in it must have following lower and upper bounds:

\[ \text{SoC}_{k} \leq \text{SoC}_{k,t} \leq \text{SoC}_{k,0} = \text{SoC}_{k,T}, \forall k, t, \]  

(5)

where \( \text{SoC}_{k} \) is upper bound and \( \text{SoC}_{k,0} \) is lower bound of the \( k \)-th ESS’ SoC. Moreover, we set the initial available SoC the same as final available SoC for a better scheduling of the \( k \)-th ESS.

4) PEV: As the trend in automobile has changed towards renewable energy sources (RES) all over the world, electric vehicles, especially PEVs, have become one of the most promising vehicles to reduce the carbon emission. It is necessary to consider the impact when a large amount of PEVs deployed in a distribution system together, which will cause a huge demand ripple that need to be mitigated through ESS and other demand response components. Thus, several PEVs are considered in the proposed model with uncertain arrival SoCs. Besides, the arrival and departure time of the PEVs are uncertain as well. Similarly, we have following charging dynamics for PEVs:

\[ \text{SoC}_{v,t} = \text{SoC}_{v,t-1} + \frac{\rho_{ch,v,t} u_{ch,v,t} I_{v,t} v_{ch,v,t}}{E_v}, \forall v, t, \]  

(7)

where \( \rho_{ch,v,t} \) is the constant charging rate. \( \eta_{ch} \) is the charging efficiency. \( E_v \) is the rated energy. We use a binary variable \( u_{ch,v,t} \) to represent charging decisions, i.e., \( u_{ch,v,t} = 1 \), the \( v \)-th PEV is being charged; when \( u_{ch,v,t} = 0 \), the \( v \)-th PEV is in an idle status.

In addition, to prolong the life time of the batteries with in the PEV, the upper and lower bounds of PEVs’ SoC are modeled as follows:

\[ \text{SoC}_{v} \leq \text{SoC}_{v,t} \leq \text{SoC}_{v,b}, \forall v, t, \]  

(8)

where upper bound \( \text{SoC}_{v} \) and lower bound \( \text{SoC}_{v,b} \) are imposed to enhance batteries’ lifetimes.

B. Comprehensive Comfort Levels

In order to quantify the satisfaction levels related to comprehensive components, the idea of comfort level is proposed. In this section, three major comfort levels are modeled: i) comfort level relate to indoor temperature; ii) comfort level related to water temperature; and iii) comfort level related to SoC, respectively.

1) Comfort Level Related to Indoor Temperature: As aforementioned, the indoor temperature directly related to the comfort levels of both the critical loads and the occupants inside the CB, which is necessary to be ensured within a predefined range. Therefore, we propose the following model to capture the unique feature of the indoor temperature in a CB with a HVAC system:

\[
J_{\text{hvac},t} = \begin{cases} 
0, & T_{\text{in}}^\text{max} \geq T_{\text{in}}^\text{max}, \ T_{\text{in}}^\text{hvac} + \epsilon_{\text{hvac}} \leq T_{\text{in}}^\text{max} \\
1 - \frac{T_{\text{in}}^\text{max} - (T_{\text{in}}^\text{hvac} + \epsilon_{\text{hvac}})}{T_{\text{in}}^\text{max} - \epsilon_{\text{hvac}}}, & T_{\text{in}}^\text{hvac} + \epsilon_{\text{hvac}} \leq T_{\text{in}}^\text{max} \leq T_{\text{in}}^\text{max} + \epsilon_{\text{hvac}} \\
1, & T_{\text{in}}^\text{min} \leq T_{\text{in}}^\text{hvac} + \epsilon_{\text{hvac}} - \epsilon_{\text{hvac}} \\
0, & T_{\text{in}}^\text{min} \leq T_{\text{in}}^\text{hvac} - \epsilon_{\text{hvac}} \\
\end{cases}
\]  

(9)

Comfort indoor temperature zone is defined as \( T_{\text{in}}^\text{hvac} \pm \epsilon_{\text{hvac}} \), where \( \epsilon_{\text{hvac}} \) is maximum indoor temperature deviation from desired temperature that can still ensure a comfort temperature zone. The most comfort level relate to the indoor temperature is 1 and the most uncomfortable level related to the indoor temperature is 0.

2) Comfort Level Related to Water Temperature: The water temperature in the water tank attached to the EWH is closely related to the occupants’ comfort level, which is relatively important to maintain above a certain threshold.

\[ T_{\text{d}}^\text{ewh} - \delta_{\text{ewh}} \leq T_{\text{d}}^\text{ewh}, \forall \tau \]  

(10)

Parameter \( T_{\text{d}}^\text{ewh} \) is the desired water temperature in the hot water tank of the EWH.

\[ J_{\text{ewh},\tau} = \begin{cases} 
1, & T_{\text{ewh}}^\text{min} \leq T_{\text{ewh}}^\text{ewh} \leq T_{\text{ewh}}^\text{max} \\
T_{\text{ewh}}^\text{ewh} - \delta_{\text{ewh}} \leq T_{\text{ewh}}^\text{ewh} \leq T_{\text{ewh}}^\text{ewh} - \delta_{\text{ewh}} \\
0, & T_{\text{ewh}}^\text{ewh} \leq T_{\text{ewh}}^\text{ewh} \leq \epsilon_{\text{ewh}} \\
\end{cases}
\]  

(11)

Parameter \( \delta_{\text{ewh}} \) is the maximum allowed temperature deviation from desired water temperature. Similarly, 1 represents the most comfort level and 0 denotes the most uncomfortable level related to water temperature.

3) Comfort Level Related to SoC: Even though SoC can be treated as comfort level relate to the energy in the batteries directly, the unique features of PEVs cannot be handle by SoC itself. Thus, the comfort level relate to SoC is proposed to capture the relationship between the desired SoC and actual SoC of each PEV.

\[ J_{v,t} = \begin{cases} 
1, & \text{SoC}_{v}^\text{u} \leq \text{SoC}_{v,t} \leq \text{SoC}_{v,b} \\
\frac{\text{SoC}_{v,t} - \text{SoC}_{v,b}^\text{base}}{\text{SoC}_{v,b}^\text{base} - \text{SoC}_{v,b}^\text{base}}, & \text{SoC}_{v,b}^\text{base} \leq \text{SoC}_{v,t} \leq \text{SoC}_{v}^\text{d} \\
0, & \text{SoC}_{v,b}^\text{base} \leq \text{SoC}_{v,t} \leq \text{SoC}_{v}^\text{d} \\
\end{cases}
\]  

(12)
\(J_{v,t}\) denotes comfort level of the \(v\)-th PEV owner. \(SoC_{v, t}^{d}\) is desired SoC for the \(v\)-th PEV. \(SoC_{v}^{\text{base}}\) represents base SoC required for the \(v\)-th PEV with round trip between the owner’s house and the CB.

C. Reliability Constraints

The reliability of the CB’s power supply is supported by both the distributed generation units within the commercial system and the upstream grid. However, both energy sources may be unreliable due to the uncertainties associated with the power generation process and the demand response signal. Therefore, it is important to ensure the reliability of the power supply of the CB through the proposed constraints.

1) Power Balance: The power supply must be the same as the power consumption, where the power supply includes: discharging of ESS, power generation of roof-top solar panel and the power delivery from the main grid; and the power consumption includes: charging of ESS, charging of PEVs, power used by HVAC system, power consumed by EWH, and the power for the base load. We denote aggregate critical power loads as \(d_t\) that must be satisfied. Therefore, we have following power balance equation:

\[
\sum_k \left( p_{k,t}^{\text{dis}} u_{k,t}^{\text{dis}} - p_{k,t}^{\text{ch}} u_{k,t}^{\text{ch}} \right) + u_{g,t} + p_{g,t} u_{g,t} = d_t + p_{t}^{\text{ewh}} \sigma_{t}^{\text{ewh}} + p_{t}^{\text{hvac}} d_t + \sum_{v} p_{v,t}^{\text{dis}} I_{v,t}^{\text{ch}}, \forall t. \tag{13}
\]

Parameter \(u_t\) denotes the output of renewables. Parameter \(p_{g,t}\) represent real-time power buy from a retail electricity market based on the uncertain real-time electricity prices. Binary variable \(u_{g,t}\) is proposed to select which power amount to buy from the main grid.

2) Uncertain Grid-Connection Condition: Additionally, we use \(T_{g}^{\text{dis}}\) to represent the capacity limit on the point of common coupling (PCC) \(g\) [8]. Moreover, in order to protect substations and transformers, the distribution system operator may perform the load curtailment in the peak hours, where the power exchange through the PCC between the CB and the main grid will be zero. Thus, a parameter \(I_{g,t} \in [0,1]\) is adopted to model the uncertain load curtailment signals from a system operator during the peak hours. Then, we have the following constraints on the grid-connected tie-line:

\[
0 \leq p_{g,t} u_{g,t} \leq T_{g} I_{g,t}, \forall g, t \in [40, 68] \\
0 \leq p_{g,t} u_{g,t} \leq T_{g} I_{g,t}, \forall g, t \notin [40, 68] \tag{14}
\]

Binary parameter \(u_{g,t}\) denotes the selection of PCC to perform the power exchange, which means the system can be extend to multiple commercial systems with different PCCs.

III. PROBLEM FORMULATION

Our objective in this paper is to develop an optimal energy management strategy for the central controller of a CB that can automatically minimize the operation and maintenance costs and maximize comprehensive comfort levels for occupants during the office hours (8:00 am–8:00 pm). Even though the CB considered is not a black box, there are still a lot of uncertainties associate with the energy management process such as base load demand, power output of renewables, electricity prices, and PEV arrival SoC and arrival/departure time.

In order to mitigate the influence of lacking comprehensive information on the distribution of aforementioned uncertainties, we model the energy management process as a Markov decision process (MDP) with the environment models formulated in Section II at each time step \(t\). In addition, we formulate the MDP model based on a tuple with four vectors: \(S, A, \mathbb{R}, S'\), where \(S\) and \(A\) denotes the state and action spaces, respectively. \(S'\) is the state space after the transition through the environment from the original state space \(S\), i.e., from time \(t\) to time \(t + 1\). \(R : S \times A \times S' \Rightarrow R\) represents the reward function, i.e., the immediate reward obtained by the central controller of the CB after taking the action \(A\) that changes from state \(S\) to state \(S'\).

In the proposed model, the states are considered to be continuous for the whole time interval, while the actions are set to be discrete, which adhere to the nature of the major components in a CB. The detailed information for the state space, action space and reward function are introduced as follows:

- **States**: The states of the DRL based energy management strategy is represented by a vector including \([T_{p}^{\text{ch}}, T_{p}^{\text{dis}}, T_{p}^{\text{ewh}}, SoC_{v,t}, SoC_{h,t}, T_{ewh}^{\text{ch}}]\), which denotes the indoor temperature, the inner wall temperature, the outer wall temperature, the SoC of PEV \(v\), the SoC of ESS \(h\), and the water temperature of the EWH, respectively. The upper and lower bounds of the each state variable are as follows: \(T_{p}^{\text{ch}} \in [18, 28]^{\circ}C\); \(SoC_{v,t}^{d} \in [6.25\%, 93.75\%]\); \(SoC_{h,t}^{d} \in [20\%, 100\%]\); and \(T_{ewh}^{\text{ch}} \in [20, 100]^{\circ}C\).

- **Actions**: The actions are taken at each time step based on the maximization purpose of the cumulated Q values. The actions of the DRL based energy management strategy is represented by a vector including \([\sigma_{t}, u_{v,t}^{\text{ch}}, u_{v,t}^{\text{dis}}, z_{ewh,t}^{\text{ch}}, u_{g,t}]\). The terms in the vector indicate the on/off status of HVAC system, the charging/idle status of the PEVs, the charging/discharging status of the ESS, the load serving condition of the EWHs, and the binary variable that select which power amount to buy from the main grid, respectively. All actions are binary variables.

- **Reward**: Our objective is to minimize operating costs of the proposed CB while maximizing occupants’ comfort levels. In order to apply the proposed DRL approach, the objective needs to be formulate into a reward function. The proposed reward function includes the electricity purchase cost, the degradation cost of ESS as penalty terms; and comfort levels related with indoor temperature, SoC of PEV and water temperature as reward terms. We formulate the proposed energy management problem over a operating day, where the comfort levels are considered during the business hours. In addition, we can combine these terms together through [9]. An unification process is coupled with comfort levels, where the total power that is required to increase all comfort levels from 0 to 1. The
detailed reward function as follows:

\[ R_t = -C_g p_{g,t} u_{g,t} - \sum_k C_{\text{ESS}} (p_{ch,k,t} u_{ch,k,t} + p_{\text{dis},k,t} u_{\text{dis},k,t}) \]

\[ + J_{\text{hvac},t} + \sum_v J_{\text{e},v,t} + J_{\text{ewh},t}. \]  

(15)

Note that \( C_g \) denotes the uncertain retail electricity prices. Parameter \( C_{\text{ESS}} \) is the degradation cost coefficient of the ESS.

The basic Q-value and the bellman equation will not be introduced in this paper due to space limitation, however, the detailed explanation can be found in [10].

IV. SIMULATION RESULTS

In this section, the proposed algorithm is evaluated through real-world datasets, where the numerical settings are the same as [9].

All simulations are implemented on a desktop computer with 3.0 GHz Intel Core i5-7400 CPU and 8GB RAM. The proposed DRL based energy management problem is simulated using Python 3.5, Gurobi 8.0 and Tensorflow 1.8.

A. Convergency of the Training Process

We first train the proposed deep Q-network (DQN) with 10,000 highly intermittent episodes for each operating day, which contains 48 time slots in business hours from 8:00 am–8:00 pm. The episodes are generated based on four uncertainties, namely: base load demand, power output of renewables, electricity prices, and PEV arrival SoC and arrival/departure time (the uncertainties related to PEVs are highly correlated). As shown in Fig. 1, the smoothed reward expends as the number of training episodes increases. Note that the “smoothed reward” represents the aggregation process of the rewards for all the time slots in one operating day. Even though the rewards tend to vary during the training process due to aforementioned uncertainties, the proposed DQN learns the optimal action pairs quickly. Moreover, as shown in Fig. 2, the smoothed losses for the whole business hours decrease as the number of training episodes expends. This demonstrates that the random selected samples are more close to the batch episodes. The convergence rate of the proposed DQN is about 2,000 episodes, where the increment of the reward and the decrement of the loss for the training process are slower than previous stages.

B. DRL versus Scenario-based Stochastic Optimization Approach

To benchmark the performance of the proposed DRL approach, we adopt a scenario-based stochastic optimization approach based energy management strategy for the same 10,000 scenarios that generated from the four uncertainties. The possibilities of the scenarios are assumed to be the same as \( 1/10000 \). The objective function for the scenario-based stochastic optimization approach is the same as the reward function for the proposed DRL approach. The models and constraints for the scenario-based stochastic optimization approach are similar to the proposed DRL approach as well. Through python and gurobi, we obtain the global optimal expected value of the objective function as 77.62 for the entire business hours. As shown in Fig. 1, the trend of rewards is approaching the expected value of the optimization approach (the red horizontal line). Thus, the simulation result validates the convergence of the proposed DRL approach after the training process to the global optimum.

C. Testing Process and Sensitivity Analysis

In addition, we test the trained DQN with two sets of 1,000 new episodes based on different uncertainty distributions of the four aforementioned uncertainties. The first set of 1,000 new episodes share the same electricity price patterns with the 10,000 training episodes. The rest three uncertainties are based on different distributions (also different from the first set). In this way, we could exam the capability of the proposed DRL approach
in handling unknown states and environments. As shown in Fig. 3 and Fig. 4, the smoothed rewards for the whole business hours accumulated the highest reward for each episode in all two sets. The simulation results can be illustrated in three folds. First, the proposed DQN is perfectly trained based on the 10,000 episodes, where the testing rewards are close to the optimum. Second, the uncertainties associate with the energy management process are highly correlated, which can be traced from one to another. Third, for a certain period, the rewards share a similar shape as shown in Fig. 4. This is reasonable since the scenarios are generated based on monthly data which is the same as the period of 30 episodes.

In summary, the simulation results ensure the reliability and effectiveness of the proposed DRL based energy management strategy.

V. CONCLUSION

In this paper, we develop a deep reinforcement learning based control strategy to determine optimal actions for major components in a commercial building to minimize operation and maintenance costs while maximizing comprehensive comfort levels of occupants. An unsupervised deep Q-network method is introduced to handle the energy management problem through evaluating the influence of operation and maintenance costs on comfort levels related environments at each time slot. An optimum control decision can be derived that targets both immediate and long-term goals, where exploration and exploitation are considered simultaneously. Extensive simulation results validate the effectiveness of the proposed deep reinforcement learning based energy management strategy.

REFERENCES