Automated Control of Transactive HVACs in Energy Distribution Systems

Boming Liu, *Member, IEEE,* Murat Akcakaya, *Member, IEEE,* Thomas E. McDermott, *Fellow, IEEE*

***Abstract*—Heating, Ventilation, and Air Conditioning (HVAC) systems contribute significantly to a building's energy consumption. In the recent years, there is an increased interest in developing transactive approaches which could enable automated and flexible scheduling of HVAC systems based on the customer demand and the electricity prices decided by the suppliers. Flexible and automated scheduling of the HVAC systems make it a prime source for participation in residential demand response or transactive energy systems. Therefore, it is of significant interest to identify an optimal strategy to control the HVAC systems. In this paper, reducing the energy cost while keeping the comfort level acceptable to the users, we argue that such a control strategy should consider both the energy cost and user comfort simultaneously. Accordingly, we develop the control strategy through the solution of an optimization problem that balances between the energy cost and consumer's dissatisfaction. This optimization enables us to solve a decision-making problem through first price prediction and then choosing HVAC temperature settings throughout the day based on the predicted price, history of the price and HVAC settings, and outside temperature. More specifically, we formulate the control design as a Markov decision process (MDP) using deep neural networks and use Deep Deterministic Policy Gradients (DDPG)-based deep reinforcement learning algorithm to find the optimal control strategy for HVAC systems that balances between electricity cost and user comfort.**

***Index Terms*—Transactive energy, reinforcement learning, HVAC.**

I. Introduction

INCREASE in population, rapid urbanization, and the usage of various household appliances leads to increasing energy consumption. It is crucial that the energy providers are reliable and flexible based on these increases in the demands. Demand response (DR) of the energy providers motivates the consumers to adapt their energy consumption in response to the market pricing signals [1]. In the recent years, with the widespread application of advanced information and communication technologies, buildings and household appliances have become more intelligent, having the potential to operate more efficiently to adjust their usage based on the DR and also to achieve higher energy savings. Transactive energy (TE) extends DR to operate on faster time scales with multilateral market participation by responsive loads [2]. In this paper, we focus on TE systems with HVAC as a responsive load.

B. Liu is with University of Pittsburgh, Pittsburgh, PA 15213 USA (e-mail: bol22@pitt.edu).

M. Akcakaya is with University of Pittsburgh, Pittsburgh, PA 15213 USA (e-mail: akcakaya@pitt.edu).

T. E. McDermott is with Pacific Northwest National Laboratory, Richland, WA 99354 USA (e-mail: thomas.mcdermott@pnnl.gov).

The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.

Electricity use by residential air conditioners accounts for 14.7% of the total power consumption in the US, which was the largest use of electricity by the U.S. residential sector in 2018 [3]. With the advancements in technology, HVAC systems can be designed to participate in TE systems with energy providers by modifying the temperature levels at each individual residence based on the consumer needs, available energy levels and energy prices. HVAC load can be shifted by pre-heating or pre-cooling the houses providing flexibility to these systems for intelligent operation based on TE [4]. However, consumers are generally willing to pay more for comfort. For example, it was shown that residential consumers will pay two times the actual price for electricity during a power outage [5]. This may be partially due to the fact that the consumers may not be aware of the price changes and/or they may not be willing to compromise on their comfort. However, another factor that contributes to this is that the current HVAC (or other household appliance) technology does not adjust energy consumption patterns that can balance between consumer comfort and energy savings. We argue that future HVAC technology should enhance an intelligent automated operation for active participation of the consumers to achieve this balance between price and comfort [6].

Real-time thermal control is required for the HVAC systems to participate in TE in an automated manner. Traditionally, model-based approaches are used for thermal control problems [7]–[9], often requiring simplified mathematical modeling of the dynamics of the HVAC systems. However, model-based approaches require time and domain expertise [10] to obtain a robust and generalized approach for HVAC thermal control strategy design due to various randomness originating from individual residences (e.g. size, thermal integrity, window wall ratio and different behaviors of the end users) which introduces additional complexity and uncertainty to the control problem.

In order to address this randomness, artificial intelligence (AI) was applied in many optimal decision-making problems in TE by imitating human behavior and automating the control of the appliances such as HVAC systems. To solve such problems, especially reinforcement learning (RL) was utilized. RL is a machine learning approach with a strong ability to learn and adapt through the interaction with the environment of real world applications. It was shown that with the help of RL, a well designed TE scheme can achieve better performance on the optimal control and decision making of residential appliances. For example, most studies demonstrated the use of a popular RL method, Q-learning [11], in DR and TE [12]–[14]. Another RL based method was proposed in [15] for the modeling and learning of TE for plug-in electric vehicle (PEV)

charging to reduce the long-term cost. Yang et al. used RL to solve the optimal control of a building energy system [16]. In [17], with the predicted future price, the authors proposed a multi-agent RL algorithm to make optimal decisions for the control of various home appliances. In [18] and [19], batch RL algorithms were proposed to schedule thermostatically controlled loads and water heaters participating in a day-ahead market. However, few of the studies modeled the appliances with a high level of detail. Most of the above mentioned approaches did not have a practical way to deal with the continuous space of the controlled state (temperature) of the HVAC systems. Moreover due to limitations in the simulations, these studies failed to provide a high degree of granularity in the precise control of the HVAC.

In this paper, we develop an RL-based approach for precise control of HVAC systems that are participating in the energy market as transactive elements in the Transactive Energy Simulation Platform (TESP) [20]. TESP was developed by Pacific Northwest National Laboratory (PNNL) as an open-source simulation platform with transactive market and control mechanisms for the grid [21]. TESP includes distribution simulator, transmission simulator and building simulator with multiple transactive agents, and the integrating Framework for Network Co-Simulation (FNCS) [22] that manages the message exchange among different simulators. In order to have an intelligent and granular control of the HVACs, we utilize RL and formulate the control problem as an optimization of cost function that balances between the electricity cost and end-user satisfaction. More specifically, combined with a price prediction method using historical data, we adopt Deep Deterministic Policy Gradients (DDPG) RL algorithm. The methods are implemented as a RL agent in TESP simulations. DDPG is a deep reinforcement learning approach developed for continuous action space; therefore it is naturally suitable for the control of HVAC systems achieving a finer and more precise control. We specifically use DDPG RL to control the base temperature schedule of the HVAC in TESP to make the TESP thermostat controller respond to the cleared market prices more intelligently at each time step to maximize the long term reward that balances between electricity cost and end-user satisfaction.

II. Method

In this section, we describe the formulation of the optimum HVAC control balancing between energy cost minimization and customer satisfaction based on RL. This RL based method relies on the predicted energy price; therefore, a price prediction method based on ANNs is also presented in this section.

A. HVAC Response and Problem Formulation

In a transactive energy system, residential users are able to participate in TE through a transactive HVAC system. Transactive HVAC systems are flexible, and they can adjust the power consumption by changing the temperature settings in residences. Here, we formulate the HVAC temperature control objective to minimize the electricity cost and the dissatisfaction of the customers caused by the temperature differences between the desired and adjusted temperature settings. We argue that the current room temperature depends on the HVAC state and power, outdoor temperature, and the room temperature of previous time step. Accordingly, different than the legacy ramp transactive control mechanism used in TESP [23], we formulate the HVAC control through a Markov Decision Process (MDP) to optimize the energy cost and customer satisfaction simultaneously. MDP is a mathematical framework that satisfies Markov property and has four elements: a set of states which represent the environment, a set of possible actions for each state, a reward function to assess the value of each action taken at a certain state, and the rules for the transitions among different states. Below is the description of the state, action, and reward function tailored to the HVAC; the control flow of HVAC based on MDP is shown in Fig. 1.

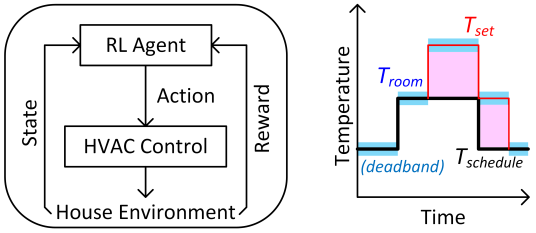


Fig. 1: HVAC control flow and settings; discomfort region shaded in pink.

a) HVAC State: The HVAC power consumption is influenced by various factors. We consider these factors as the elements of the HVAC state in the MDP model. We denote the HVAC state at time t as S_t, see (1). The observable state of a HVAC should contain information about both indoor and outdoor environment as they significantly affect the energy consumption. Therefore, the indoor temperature T^t_{room} and outside temperature T^t_{out} at time t are considered as elements of the HVAC state. In addition, the desired or scheduled base temperature, $T^t_{schedule}$, of the house is included in the HVAC state. Finally, since the HVAC on/off status at time t depends on price-responsive T^t_{set} and the current indoor thermal environment, T^t_{set} is also included in the HVAC state.

$$S_t = \{T^t_{set}, T^t_{room}, T^t_{schedule}, T_{out}\} \tag{1}$$

$$T^t_{set} = T^t_{schedule} + \frac{(P^t_{cleared} - P_{average}) \times \left|T_{\max/\min}\right|}{k_{high/low} \times \sigma_{actual}} \tag{2}$$

The relation between T^t_{set} and $T^t_{schedule}$ in TESP is shown in (2), where $P_{average}$ is the historical mean price, $\left|T_{\max/\min}\right|$ is the allowed range of set point variation , $k_{high/low}$ is the bidding ramp denominator, σ_{actual} is standard deviation of the price. Bidding ramps and allowed temperature ranges could be unequal above and below $T^t_{schedule}$ as in [23].

b) Action: The aim of the HVAC control is to minimize the cost by changing the HVAC temperature setting schedule, $T^t_{schedule}$. Therefore, in our formulation, the learning agent of the RL approach based on MDP assumptions is designed to make changes in the scheduled temperature deviating from the original schedule based on a reward function. The action is the temperature change from the original schedule in a certain adjustable range, e.g. [-5,5] degrees Fahrenheit.

c) Reward: The reward of each action consists of two parts, the penalty for the energy consumed by the HVAC during the time period and the discomfort of the consumer resulting from the control action taken at a given state. The discomfort is the estimated feedback of the occupants'

dissatisfaction under the current thermal condition. The reward at each time step is defined as:

$$r_t = -\alpha(E^t_{hvac} \times P^t_{clear}) - (1-\alpha)k \times (T^t_{dev})^2 \quad (3)$$

$$T^t_{dev} = (T^t_{room} - T^t_{schedule}) \quad (4)$$

where α represents the importance of the cost of energy consumption of the HVAC. E^t_{hvac} is the energy consumption of the HVAC during this time step. P^t_{clear} is the cleared price from TESP. The cost will be higher if more energy is consumed when the price is relatively high. The second term is the consumers' dissatisfaction cost which is calculated by multiplying a factor k by the squared room temperature deviation T^t_{dev} from the original schedule temperature.

B. HVAC Control through Deep Deterministic Policy Gradient

Model-based or model-free approaches can be used in reinforcement learning to optimize energy cost and/or thermal comfort through the control of HVAC [24]. Model-based approaches require complete information of the HVAC thermal dynamics to represent transition among different states. For example, for the model-based approaches, accurate dynamic interactions between the residence and the surrounding environment may be needed. In contrast, model-free methods are more flexible to overcome the detailed modeling of the HVAC dynamics and accordingly to represent state transitions.

Q-learning, state-action-reward-state-action (SARSA) and deep Q-networks (DQN) are commonly used for model-free RL [25]. However, they cannot be used to solve control problems with both continuous state and action spaces. For instance, in order to utilize DQN for HVAC control, temperature of the HVAC can be discretized finely, resulting in a large number of possible actions. But higher granularity of the action space will decrease the training efficiency dramatically. DDPG is a deep reinforcement learning method which is capable of handling a space of continuous states and actions. There exist other off-policy algorithms like soft actor critic (SAC) [26] and twin delayed DDPG [27] which are variations of the DDPG algorithm. They can also be used to solve the continuous control problem such as HVAC control. In this paper, we utilize DDPG for the control purposes as we can show through our numerical results that the reward convergence is robust to the changes in the hyperparameters.

As shown in Fig. 2(a), for any given input state, through

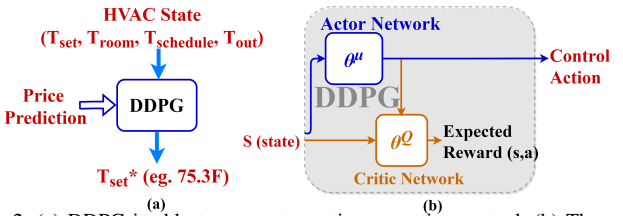


Fig. 2: (a) DDPG is able to generate continuous action control. (b) The network structure of DDPG implemented as the RL agent.

the interaction of actor and critic networks, DDPG is able to generate optimal control action directly rather than by fine discretization of the action space. The network structure of the DDPG method is presented in Fig. 2(b). More specifically, DDPG is implemented here through an actor-critic architecture that learns approximations to both policy function, θ^μ, and value function, θ^Q. An actor is used to tune the parameter θ^μ for the policy function (i.e., to decide the best control action A_t given a specific HVAC state S_t, where θ^μ represents the weights of the actor neural network). On the other hand, a critic network is used for evaluating the policy function estimated by the actor network. Here, the critic network's parameters are denoted by θ^Q. Critic network estimates the action value Q which is the expected reward of taking the control action A_t at state S_t.

The actor network and the critic network are trained through the TESP simulations which enables evaluation of different actions for different HVAC states. After training, during testing, through the interaction between actor and critic networks RL-based control outputs an optimum action that is used by TESP to control the HVAC. The training details of the actor and critic networks are provided in Algorithm 1 and Fig. 3.

Algorithm 1 DDPG

procedure DDPG RL(θ^μ,θ^Q)
  Initialize memory M of size N;
  Initialize the actor network $\mu(S_t|\theta^\mu)$ and critic network $(S_t, A_t|\theta^Q)$ with random parameter θ^μ and θ^Q
  Initialize the target network μ' and Q' with $\theta^{\mu'} \leftarrow \theta^\mu$,$\theta^{Q'} \leftarrow \theta^Q$;
  Input the estimated price $\{\hat{P}_{clear}\}_0^T$;
  Define $s_t = \{T^t_{set}, T^t_{room}, T^t_{out}, {T_{schedule}}^t\}$;
  Receive the initial HVAC state s_0=$\{T^0_{set}, T^0_{room}, T^0_{out}, {T_{schedule}}^0\}$;
  for t=0,1,2,..,T **do**
    Select a_t by $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$;
    Execute a_t on HVAC and obtain the reward $r(s_t, a_t)$ and next state s_{t+1};
    Store the transition (s_t, a_t, r_t, s_{t+1}) in M;
    Sample K transition from M randomly and calculate the estimated policy value for the sampled transitions i : $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|{\theta^\mu}')|{\theta^Q}')$;
    Update the critic network θ^Q by the gradient $\nabla_{\theta^Q} L$ of the MSE over the K size mini-batch and learning rate β_y: $\nabla_{\theta^Q} L = \frac{1}{K}\sum_{i=1}^{k}(y_i - Q(s_i, a_i|\theta^Q))^2$;
    Update the actor network using the sampled policy gradient $\nabla_{\theta^\mu} J$ and learning rate β_x: $\nabla_{\theta^\mu} J \approx \frac{1}{K}\sum_{i=1}^{K} \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i}$
    Update the target networks (τ : updating rate):
    ${\theta^\mu}' \leftarrow \tau\theta^\mu + (1-\tau){\theta^\mu}'$;
    ${\theta^Q}' \leftarrow \tau\theta^Q + (1-\tau){\theta^Q}'$;
  end for
end procedure

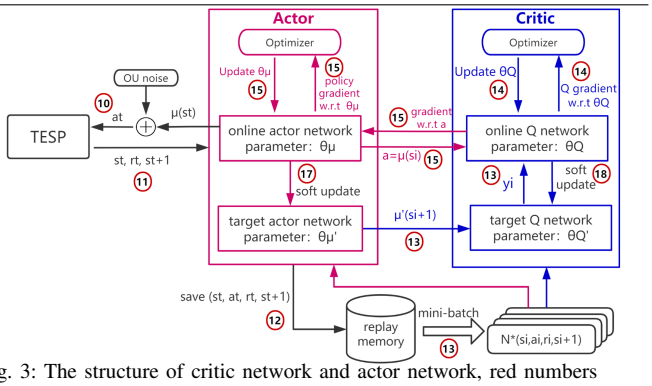


Fig. 3: The structure of critic network and actor network, red numbers correspond to lines in Algorithm 1.

For the training, we initialize the actor network and the critic network with random parameters, also we use the same random parameters to initialize the target actor network and target critic network. DDPG enables the agent to explore a wide variety of actions in the beginning of learning. Specifically, after receiving the initial state s_0, the actor network explores the action space to select a control action. We add

a random noise to the selected action to explore the control action space to prevent converging to a local solution through Ornstein–Uhlenbeck process [28], see Algorithm 1 line 10.

During training, at each time step t, after the learning agent takes the control action a_t, it communicates this action to TESP to change the HVAC state s_t, then receives the new HVAC state s_{t+1} and the reward R_t calculated based on (3) as feedback from TESP. In order to improve the convergence and decrease the correlation among the training samples, we add a memory buffer for experience replay. So at every time step, the state action transition s_t, a_t, R_t, s_{t+1} is stored into the memory M. From the memory M, we then randomly sample K transitions and calculate the estimated value y of each sampled transition using the target networks. The next-state Q values are calculated with the target value network and target policy network (Fig. 3 arrow 13). Then, we minimize the mean-squared loss between the updated Q value and the original Q value (line 14). Here, we use the target networks which are constrained to change slowly. The two target networks $\theta^{\mu'}$ and $\theta^{Q'}$ will slowly track two learned networks θ^{μ} and θ^{Q} which will help improve the stability of learning. Calculation of the estimated value y through the target networks is achieved through Algotithm 1 line 13, where γ is the discounting factor indicating the importance of future versus current value. The weight of the critic network is updated by minimizing the mean square error with respect to the critic network parameters using the values corresponding to the randomly selected K samples as shown in line 14 of the Algorithm 1. The policy loss is the derivative of the objective function with respect to the policy (actor network) parameters. Then the actor network is updated through the sampled policy gradient as shown in line 15 of Algorithm 1 [29]. Note that the chain rule is applied since the policy function and the actor network are both differentiable. Finally, both target networks are updated with an update rate $\tau \ll 1$ as shown in lines 17 and 18.

The actor network and the critic network of DDPG algorithm both have 2 hidden layers. The structure and different activation functions are shown in Fig. 4.

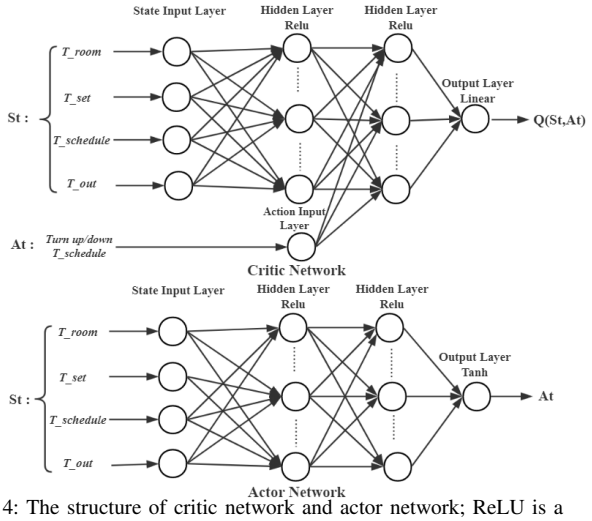


Fig. 4: The structure of critic network and actor network; ReLU is a rectified linear activation unit.

C. Price Prediction with ANN

The optimal control strategy based on DDPG relies also on the predicted electricity price, see Fig. 2. In our approach, we utilized a multi-layer perceptron neural network with 2 hidden layers to predict the future electricity price. Through such an artificial neural network, we develop a nonlinear relationship between the input variables (e.g. temperature, system load, day of the week) and the predicted output electricity price. Fig. 5 demonstrates the topology of the utilized neural network. As listed in Fig. 5 there are up to 18 day, hour, load, temperature and price inputs connecting to the hidden layers.

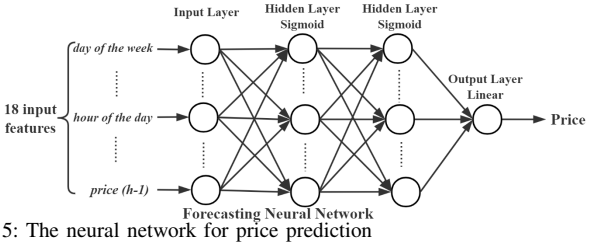


Fig. 5: The neural network for price prediction

III. Simulations and Performance Evaluation

In this section, we describe the simulation scenarios and present the numerical results. We first present the performance of the proposed ANN structure for electricity price prediction and we compare it with the state-of-the-art price prediction methods such as weighted average filter [30], support vector machine (SVM)-based prediction [31], and ANN-based prediction [32]. Then, we consider different simulation scenarios in TESP to compare the proposed DDPG RL-based HVAC control strategy with the control strategy that is already implemented in TESP in terms of electricity cost and consumer satisfaction. We represent the consumer satisfaction as the deviation of the temperature settings from the desired temperature schedule of the HVAC systems.

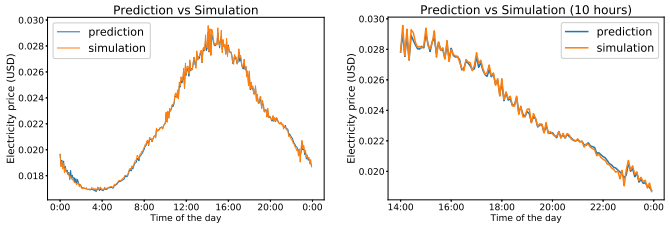
A. Simulation of ANN price prediction

1) Simulation scenarios: We generated four weeks of electricity price data using a TESP feeder model with a substation and 306 different houses with HVACs. We used the first two weeks of the generated price data for training the proposed neural network and used the second two weeks of data for testing. As shown in Table I, we considered up to 18 input features to train the proposed neural network to predict electricity price. Day of the week, hour of the day and historical price data are obtained directly from the generated TESP data. Historical weather (temperature) and the load data for price prediction training in the Pittsburgh area are obtained from the weather data in Typical Meteorological Year 3 (TMY3) format [33] and PJM website, respectively. PJM is a regional transmission organization and they publish historical hourly load data for Duquesne Light Company on their website. Since the TESP simulation data have higher temporal resolution compared to the load data, the hourly load data is interpolated to obtain 5 minutes per sample temporal resolution.

TABLE I
Input features for price prediction (h represent hour)

Input Features	
Day of the week	1-7
Hour of the day	1-24
Historical price	(h-1),(h-2),(h-3),(h-24),(h-25),(h-26),(h-48),(h-168)
PJM load	(h-1),(h-2),(h-3),(h-24),(h-25),(h-26)
Weather	temperature
Price distribution	mean of the distribution

2) Price Prediction Simulation Result : In Fig. 6 we compare the proposed neural network that is trained using all 18 inputs that are listed in Table I directly with the TESP simulation results. Here TESP simulation results are the benchmark. Fig. 6 (a) shows 24 hour prediction results with mean square error (MSE) of 2.12×10^{-4}), and Fig. 6 (b) shows 10 hour simulation results with MSE 2.59×10^{-4}). From these two figures we observe that the overall trend of the predicted price is consistent with the TESP-simulated electricity price. Note that even some small fluctuations in price are also correctly predicted.



(a) 24 hour prediction result (b) 10 hour prediction result

Fig. 6: Price prediction vs TESP Simulation data

Here we also compare the proposed approach with the state-of-the-art price prediction methods. We denote the proposed approach as ANN with weather and price distribution input (ANN + weather + price distribution) and compare it with weighted average filter-based, SVM-based, ANN with weather input (ANN + weather) and ANN without weather and price distribution (ANN) methods. For this comparison, we generated simulation data from the 306-house system described above. Similar to the above scenario, historical weather and PJM load data are obtained from online sources.

The data was divided into 50 week-long periods, and the mean square error of predicting price of different weeks throughout the year is shown for different methods in Fig. 7. We observe that SVM-based method is better than the

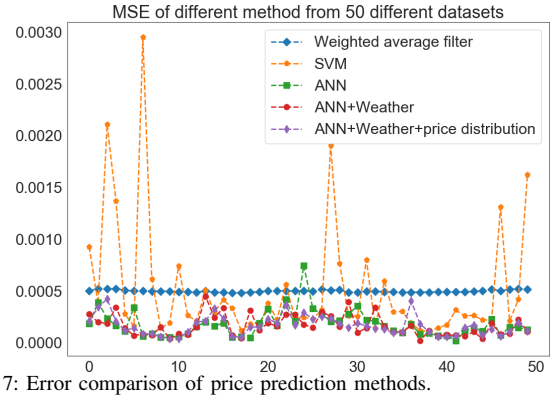


Fig. 7: Error comparison of price prediction methods.

weighted average filter, and ANN based methods outperform both the weighted average filter and SVM-based methods. To statistically compare the methods, we apply non-parametric one-sided rank sum test and the results are presented in Table II. In this table, we specifically present the p-values for testing if the methods listed in the columns have lower mean-square error in price prediction than the methods listed in the rows. A p-value lower than 0.05 means that the method listed in the column has statistically lower mean-square error compared to the method listed in the row. Similar to Fig. 7, ANN-based methods are significantly better than weighted average filter and SVM-based methods. Even though there are not statistically significant differences among ANN, ANN+W and ANN+W+P (see Table II), adding weather and price distribution information may make the price prediction more robust, see Fig. 7. But this robustness comes with price of additional data collection.

TABLE II

p value of Wilcoxon rank-sum test between the errors in Fig. 7 of the method in each row and the method in each column.

p	filter	SVM	ANN	ANN+W	ANN+W+P
filter	0.50	5.58e-5	6.75e-17	3.53e-18	3.53e-18
SVM	0.99	0.50	1.38e-7	3.31e-8	1.85e-8
ANN	1	1	0.50	0.35	0.45
ANN+W	1	1	0.65	0.50	0.58
ANN+W+P	1	1	0.55	0.42	0.50

B. Simulation of DDPG RL HVAC Control

1) Simulation scenarios: The proposed RL-based HVAC control is evaluated using TESP-simulated data on 306 houses. We specifically considered the scenarios in which HVACs are in the cooling mode. To make sure the HVACs are in cooling mode during the training, TMY3 data for Florida instead of Pittsburgh were used during the period from June to November of 2018. One generic control policy for different houses is obtained after training. The DDPG algorithm is implemented with Pytorch [34], an open source Python-based scientific computing package for machine learning. The training data comes from simulation of 212 days in TESP.

As also mentioned above, we compare the RL-based approach with the HVAC ramp control approach that is implemented in TESP. This method (which we denote as "without RL agent" in this paper) controls the HVAC using a pre-defined temperature schedule. On the other hand, the proposed RL-based method (which we denote as "with RL agent") changes the pre-defined temperature schedule based on the predicted price and DDPG-based control. We compare these two control approaches not only under normal conditions but also during a high price scenario that includes a bulk system generator outage. Test cases are illustrated in Fig. 8. Simulation configurations and key parameters of the DDPG training algorithm are listed in Table III.

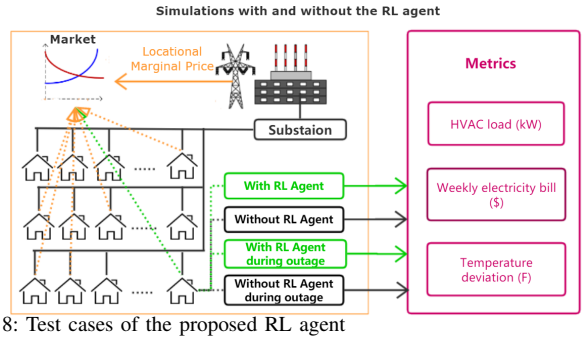


Fig. 8: Test cases of the proposed RL agent

TABLE III

Parameter Settings

T_{test}	212 days	training data time length
Δt	5 minutes	time step
β_Q	0.000025	learning rate of the critic network
β_μ	0.00025	learning rate of the actor network
τ	0.001	model update parameter
γ	0.95	reward discount factor
α	0.1∼0.5	weight factor of the electricity cost

The batch size for DDPG training is chosen to be 72. The parameter α that was introduced in (3) and that balances

between energy cost saving and customer satisfaction is varied between 0.1 and 0.5.

2) Performance Metrics: In order to compare the control methods with and without RL agents, we define electricity cost saving factor (CSF) and thermal comfort improvement factor (TIF) as the performance metrics. Both are affected by α.

$$CSF = \frac{weeklybill_{base} - weeklybill_{RL}}{weeklybill_{base}} \times 100\% \quad (5)$$

$$TIF = \frac{\Delta T_{base} - \Delta T_{RL}}{\Delta T_{base}} \times 100\% \quad (6)$$

Both CSF and TIF can be greater or less than 0; a positive CSF or TIF indicates better performance with the RL agent.

3) Simulation Result:

a) Convergence of the training process: As shown in Fig. 9, we plot the reward function as a function of training time steps for different hyper-parameters. It can be observed that the training of the algorithm is very robust to the changes in α and β_μ (β_μ=10β_Q), and τ (when τ is greater than 0.001, which is the literature recommended value).

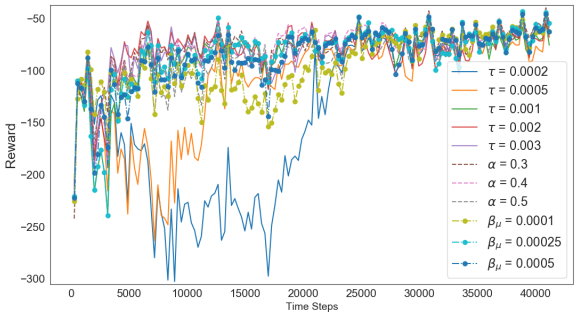


Fig. 9: Convergence of DDPG training process with different hyper parameters

b) Performance of the DDPG RL Algorithm: Through TESP simulations as described above, we compare HVAC control with RL agent to HVAC control without RL agent. Recall here that HVAC control without RL agent uses a fixed temperature schedule and adjusts the HVAC setting based on this fixed temperature schedule and cleared market price for electricity [23]. On the other hand, HVAC control with RL agent changes the temperature schedule and then adjusts the HVAC setting for price. For these two approaches, in Fig. 10, we plot the temperature schedules, HVAC temperatures and

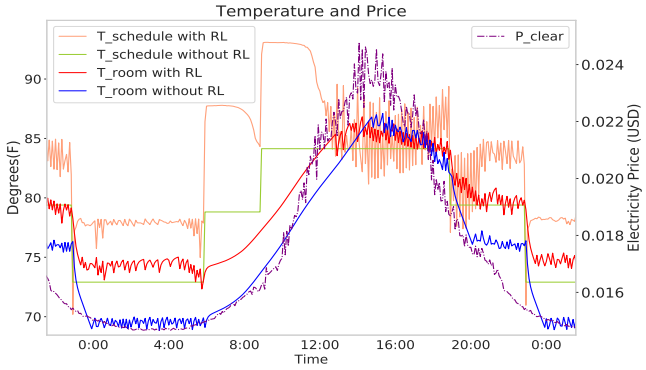


Fig. 10: Room temperature with and without RL agent

cleared market price for electricity. The purple dashed-line demonstrates the cleared market price, green and orange lines show the fixed and changing temperature schedules, respectively, and red and blue lines are for the HVAC temperatures with and without RL-agent control, respectively. The room temperature is rising along with the increasing outside temperature from 8am to noon and it continues to rise until it triggers the HVAC to cool the room. The HVAC control with RL agent predicts the afternoon increase in the cleared market price; therefore, there is sudden drop in the temperature schedule at 12:00 with RL agent, and the temperature schedule then continues to drop below the current room temperature. As a result, the HVAC starts cooling the house a little earlier than the original control without RL agent before the price peak is reached at around 14:00. Specifically, just after 12:00, the red line starts to drop and fluctuate, before the blue line. At every time step, the RL agent controls the HVAC output to minimize the deviation of the room temperature from the original schedule while aiming to consume more power for HVAC at relatively lower price.

Fig. 11 demonstrates the energy consumption of a single HVAC controlled with (blue) and without (orange) the RL agent. We observe that compared to the HVAC controlled

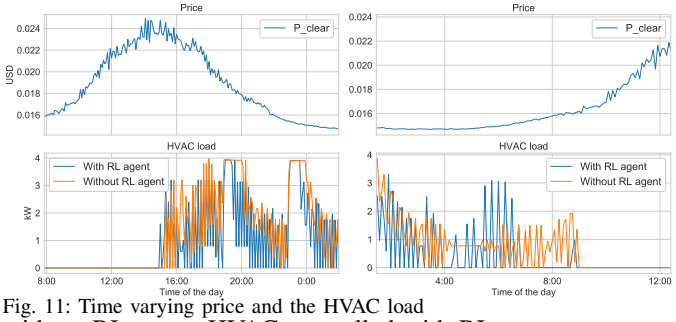


Fig. 11: Time varying price and the HVAC load

without RL agent, HVAC controlled with RL agent consumes more power (higher HVAC load) when the price is low and less power (lower HVAC load) when the price is high. Additionally, Fig. 12 shows the aggregated loads of 306 HVACs controlled with (blue) and without (orange) the RL agent. In this figure,

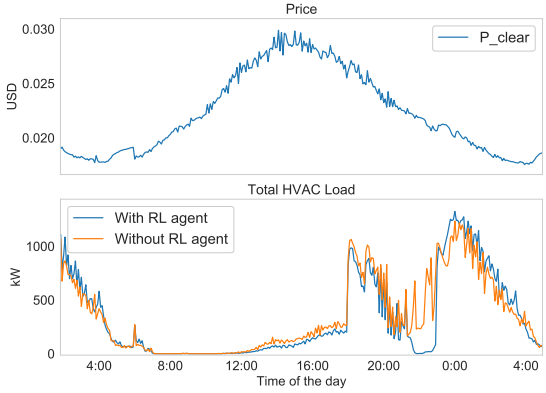


Fig. 12: Aggregate responses of the HVACs

total HVAC load is less when the price is high around 14:00 to 18:00. The HVACs consume a little more power during the time when price is relatively low such as 0:00 to 4:00. Similar to what we observe in Fig. 11, HVACs controlled with RL agents aim to save more energy at higher market prices.

Recall from (3) that α value is chosen to balance between the consumed energy cost of HVAC and the comfort level of the customers. We define the minimization of customer discomfort as the minimization of the deviation of the temperature schedule from the original schedule. Parameter α takes values between 0 and 1 and as its value increases, customers care more about the energy cost. Here we compare again

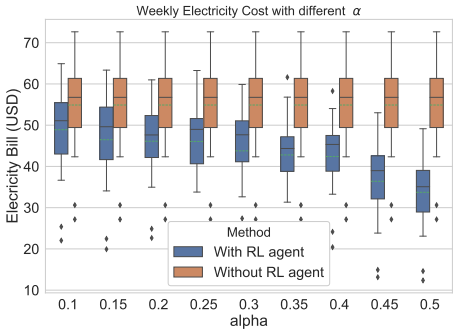


Fig. 13: Comparative box plot of the weekly energy cost vs. α

the HVAC control methods with (blue) and without (orange) RL agent for different α values. More specifically, Fig. 13 is the box plot of weekly cost of consumed power by HVAC. The green dashed lines show the weekly mean values. As can be observed from this figure the RL agent saves more money compared to the control without RL agent; saving increases as α increases. For example, the CSF is 38.5% when α is 0.5. On the other hand, Fig. 14 is the bar plot of the room temperature deviation from the desired temperature schedule. The average temperature deviation increases with

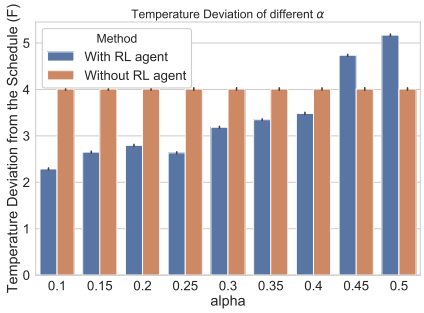


Fig. 14: Comparative bar plot of the temperature deviation vs. α

the increase of α, such that TIF ranges from 42.75% to -28.7%. Fig. 15 shows the room temperature under the control of RL agent with different α. When α increases the deviation

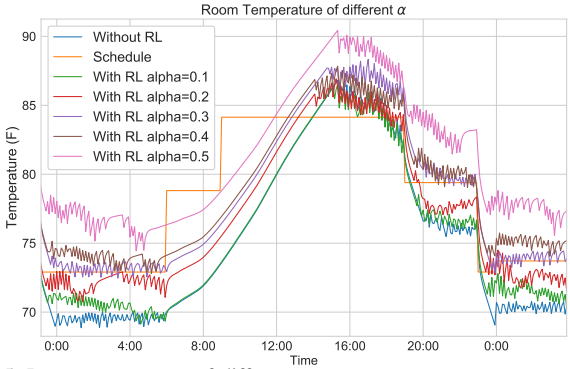


Fig. 15: Room temperature of different cases.

of the temperature from the scheduled temperature increases. With larger α, the RL agent is very sensitive to the electricity consumption, in consequence, the RL agent tends to save more energy by sacrificing thermal comfort. For example, as shown in Fig. 14, with α = 0.45 and α = 0.5, the TIF became negative and the temperature deviation is even higher than the case that uses HVAC control without RL. Moreover, Fig. 13 and Fig. 14 also demonstrate that with certain α, the RL agent is able to reduce the energy cost and improve the occupants' comfort at the same time compared to the HVAC control without RL, for example see $\alpha = 0.4$.

c) DDPG RL Performance During a Generation Outage: In the above simulations, the clearing price is at a normal level for most of the time. To evaluate the performance of the HVAC control with RL agent during high-price events, we perform simulations with a bulk generation outage at a certain time of day. We are using the same simulation scenario with 306 HVACs as described above but now there is a generation unit outage from 12:00 to 18:00.

Due to the outage of a main generation unit and the higher cost of the back up generation unit, the Locational Marginal Price (LMP) at the substation bus becomes higher than normal during the outage, leading to a high clearing price as shown in Fig. 16. With RL agent, the HVAC consumes less power

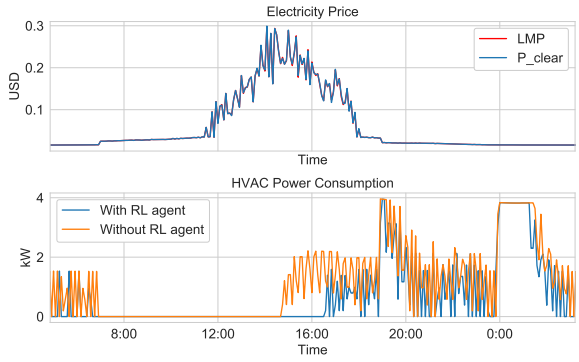


Fig. 16: Electricity price and the HVAC power consumption with RL agent during generation outage

during the outage when the electricity price is at peak. As illustrated in Fig. 16, different from the HVAC control without RL which consumes power during the price peak, the HVAC with RL agent is off beginning around 16:00 and starts to work again when the price drops.

Similar to Fig. 13 for different α values, Fig. 17 shows the box plot of weekly HVAC energy cost with generation outage. As observed in Fig. 17, without RL agent, the energy cost of

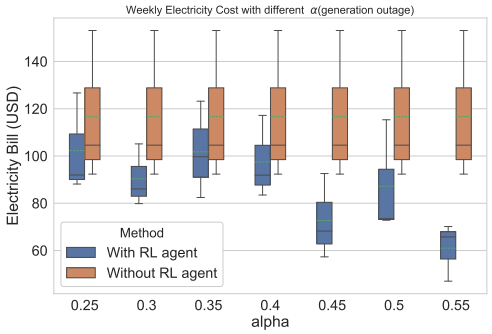


Fig. 17: Weekly HVAC energy cost vs. α with generation outage.

HVAC is doubled over the normal scenario as demonstrated in Fig. 13. Similar to Fig. 14, Fig. 18 shows the bar plot of room temperature deviation from the desired temperature schedule. When $\alpha = 0.25$, the thermal comfort with and without RL are almost the same in these two cases. Note that when $\alpha = 0.25$ the consumers are still able to save 12.7% of HVAC energy cost on average with the RL agent. That is, while the comfort level is preserved, there is more energy savings with HVAC control with RL. When $\alpha = 0.55$, although the average HVAC energy cost is reduced by 50%, the temperature deviation

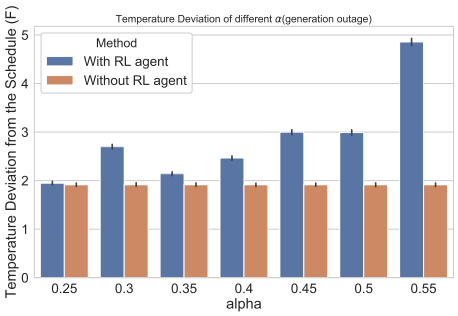


Fig. 18: Average temperature deviation vs. α with generation outage.

increases a lot. In general, consumers are able to save a greater extent of money with a higher α (emphasis on energy saving), reducing thermal comfort as shown in Fig. 18.

IV. Conclusion

We have developed a RL-based method for the control of transactive HVACs in distribution systems that are participating in a double auction electricity market. The method is integrated in and tested through TESP simulation. As the first step, we proposed and tested an electricity price prediction method and compared it with existing state-of-the-art price prediction methods. Then we used the developed price prediction method together with a DDPG approach to train a reinforcement learning agent to control the HVAC. The proposed RL-based method balances between electricity cost and customer comfort. Accordingly, we compared our approach with the ramp control method that already exists in TESP. Our results showed that the proposed method not only saves the electricity cost but also improves the customers' comfort at the same time. Our future work will explore using α as a customer-oriented slider setting to express preferences, training on the fly for continuous improvement of the local RL agent, and extensions to water heaters and batteries.

References

[1] P. Siano, "Demand response and smart grids—a survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461 – 478, 2014.

[2] Z. Liu, Q. Wu, S. Huang, and H. Zhao, "Transactive energy: A review of state of the art and implementation," in *2017 IEEE Manchester PowerTech*, pp. 1–6, 2017.

[3] "U.s. energy information administration." Accessed: 2019-11-21.

[4] J. E. Braun, "Load control using building thermal mass," *J. Sol. Energy Eng.*, vol. 125, no. 3, pp. 292–301, 2003.

[5] P. Centolella, M. Farber-DeAnda, L. Greening, and T. Kim, "Estimates of the value of uninterrupted service for the mid-west independent system operator (miso)," *Research Paper, Harvard Electricity Policy Group, Harvard Kennedy School of Government, Cambridge, Massachusetts, USA*, 2010.

[6] J. Wang, H. Zhong, Z. Ma, Q. Xia, and C. Kang, "Review and prospect of integrated demand response in the multi-energy system," *Applied Energy*, vol. 202, pp. 772 – 782, 2017.

[7] A. Molderink, V. Bakker, M. G. Bosman, J. L. Hurink, and G. J. Smit, "Management and control of domestic smart grid technology," *IEEE transactions on Smart Grid*, vol. 1, no. 2, pp. 109–119, 2010.

[8] L. Yu, T. Jiang, and Y. Zou, "Online energy management for a sustainable smart home with an hvac load and random occupancy," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1646–1659, 2017.

[9] Y. Ma, F. Borrelli, B. Hencey, B. Coffey, S. Bengea, and P. Haves, "Model predictive control for the operation of building cooling systems," *IEEE Transactions on control systems technology*, vol. 20, no. 3, pp. 796–803, 2011.

[10] J. Drgoňa, D. Picard, and L. Helsen, "Cloud-based implementation of white-box model predictive control for a geotabs office building: A field test demonstration," *Journal of Process Control*, vol. 88, pp. 63 – 77, 2020.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[12] S. Liu and G. P. Henze, "Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory," *Journal of solar energy engineering*, vol. 129, no. 2, pp. 215–225, 2007.

[13] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra, "Residential demand response using reinforcement learning," in *2010 First IEEE International Conference on Smart Grid Communications*, pp. 409–414, IEEE, 2010.

[14] S. Yousefi, M. P. Moghaddam, and V. J. Majd, "Optimal real time pricing in an agent-based retail market using a comprehensive demand response model," *Energy*, vol. 36, no. 9, pp. 5716–5727, 2011.

[15] A. Chiş, J. Lundén, and V. Koivunen, "Reinforcement learning-based plug-in electric vehicle charging with forecasted price," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 3674–3684, May 2017.

[16] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, "Reinforcement learning for optimal control of low exergy buildings," *Applied Energy*, vol. 156, pp. 577 – 586, 2015.

[17] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Transactions on Smart Grid*, vol. 10, pp. 6629–6639, Nov 2019.

[18] F. Ruelens, B. J. Claessens, S. Quaiyum, B. De Schutter, R. Babuška, and R. Belmans, "Reinforcement learning applied to an electric water heater: From theory to practice," *IEEE Transactions on Smart Grid*, vol. 9, pp. 3792–3800, July 2018.

[19] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 8, pp. 2149–2159, Sep. 2017.

[20] D. Holmberg, M. Burns, S. Bushby, A. Gopstein, T. McDermott, Y. Tang, Q. Huang, A. Pratt, M. Ruth, F. Ding, *et al.*, "NIST transactive energy modeling and simulation challenge phase ii final report," *NIST Special Publication*, vol. 1900, p. 603, 2019.

[21] Q. Huang, T. E. McDermott, Y. Tang, A. Makhmalbaf, D. J. Hammerstrom, A. R. Fisher, L. D. Marinovici, and T. Hardy, "Simulation-based valuation of transactive energy systems," *IEEE Transactions on Power Systems*, vol. 34, pp. 4138–4147, Sep. 2019.

[22] S. Ciraci, J. Daily, J. Fuller, A. Fisher, L. Marinovici, and K. Agarwal, "Fncs: a framework for power system and communication networks co-simulation," in *Proceedings of the symposium on theory of modeling & simulation-DEVS integrative*, p. 36, Society for Computer Simulation International, 2014.

[23] J. C. Fuller, K. P. Schneider, and D. Chassin, "Analysis of residential demand response and double-auction markets," in *2011 IEEE power and energy society general meeting*, pp. 1–7, IEEE, 2011.

[24] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied energy*, vol. 235, pp. 1072–1089, 2019.

[25] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*, vol. 37. University of Cambridge, Department of Engineering Cambridge, England, 1994.

[26] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.

[27] S. Fujimoto, H. Van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *arXiv preprint arXiv:1802.09477*, 2018.

[28] P. Lánskỳ and L. Sacerdote, "The ornstein–uhlenbeck neuronal model with signal-dependent noise," *Physics Letters A*, vol. 285, no. 3-4, pp. 132–140, 2001.

[29] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," 2014.

[30] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Optimal residential load control with price prediction in real-time electricity pricing environments," *IEEE transactions on Smart Grid*, vol. 1, no. 2, pp. 120–133, 2010.

[31] X. Yan and N. A. Chowdhury, "Mid-term electricity market clearing price forecasting: A multiple svm approach," *International Journal of Electrical Power & Energy Systems*, vol. 58, pp. 206–214, 2014.

[32] R. Lu, S. H. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6629–6639, 2019.

[33] S. Wilcox and W. Marion, *Users manual for TMY3 data sets*. National Renewable Energy Laboratory Golden, CO, 2008.

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.