Realistic Peer-to-Peer Energy Trading Model for Microgrids using Deep Reinforcement Learning

1st Tianyi Chen
Division of Systems, Power & Energy
University of Glasgow
Glasgow, United Kingdom
2128180c@student.gla.ac.uk

2nd Shengrong Bu
Division of Systems, Power & Energy
University of Glasgow
Glasgow, United Kingdom
Shengrong.Bu@glasgow.ac.uk

***Abstract*—In this paper, we integrate deep reinforcement learning with our realistic peer-to-peer (P2P) energy trading model to address a decision-making problem for microgrids (MGs) in the local energy market. First, an hour-ahead P2P energy trading model with a set of critical physical constraints is formed. Then, the decision-making process of energy trading is built as a Markov decision process, which is used to find the optimal strategies for MGs using a deep reinforcement learning (DRL) algorithm. Specifically, a modified deep Q-network (DQN) algorithm helps the MGs to utilise their resources and make better strategies. Finally, we choose several real-world electricity data sets to perform the simulations. The DQN-based energy trading strategies improve the utilities of the MGs and significantly reduce the power plant schedule with a virtual penalty function. Moreover, the model can determine the best battery for the selected MG. The results show that this P2P energy trading model can be applied to real-world situations.**

***Index Terms*—deep Q-network, deep reinforcement learning, P2P energy trading, smart grids**

I. Introduction

Renewable energy resources have been exploited to solve the foreseeable fossil fuel shortage problem in the past decade. Although renewable energy is sustainable, it brings significant challenges to the stability and operational safety of a large power network due to its intermittent and location-variant nature. As a result, microgrids have been proposed to address these challenges by coordinating the control of distributed energy resources (DER), local active loads and energy storage systems (ESSs) within certain regions. Within a microgrid, the distributed renewable energy sources, such as wind power and solar energy, can switch traditional energy consumers to prosumers. Multiple microgrids located in a large area can be networked to improve the efficiency and reliability of the distribution network further. However, since the installed DERs in microgrids belong to different owners, it is not realistic to directly control or operate them by a central authority. Recently, peel-to-peel (P2P) energy trading has emerged as a novel paradigm for decentralised energy market designs. P2P energy trading allows the end-users to join the trading without a central authority unit [1].

Some P2P energy trading models have been proposed to solve the renewable energy dilemma, e.g., game-theoretic approaches [2]–[6], and contract networks for P2P energy trading [7], [8]. However, making decisions based on the massive amount of data and unpredictable renewable generation in P2P energy trading by using conventional optimised techniques is problematic. DRL techniques, combined with deep neural networks and reinforcement learning (RL) techniques, could be powerful tools for addressing such P2P energy trading issues since they can solve the decision-making problems by learning from the high-dimensional historical data.

DRL/RL have been used in the area of smart grids to optimise the operation of MGs [9], energy management [10] and storage planning [11]. There is also some recent research using DRL for P2P energy trading, where a large amount of uncertainty data can be directly learned by DRL to make the decisions in the real world. For example, a local energy trading problem for prosumers was formulated as an MDP and was solved by using deep Q-learning to maximise prosumers daily economic benefit [12]. A DQN-based MG trading game was formulated to improve the utility of the MG without knowing information about other MGs [13]. However, the physical constraints in a distributed renewable energy system were not considered in these papers, and their study was limited to a typical day of the P2P energy trading, where in reality the trading behaviours change throughout the year.

In this paper, we formulate a realistic energy trading model for MGs with a set of critical physical constraints. An MG needs to make a trading strategy and negotiate with other MGs only based on its generation, demand and energy storage level. The physical constraints like transmission losses and power limits at some nodes of the system may affect the strategy of an MG. We also set a flexible utility function for each MG to evaluate its strategy, which consists of not only trading profits but also the battery wear cost, demand penalty, and optional social factors. Deep reinforcement learning is used to train the agent as an MG to derive better strategies based on the states and the utility function. Using DQN and an experience replay mechanism [14], the algorithm can speed the Q-learning rate and update the loss function with continuously collected new states and rewards instead of updating the model at the end of each episode. Last but not least, we choose one-year real-

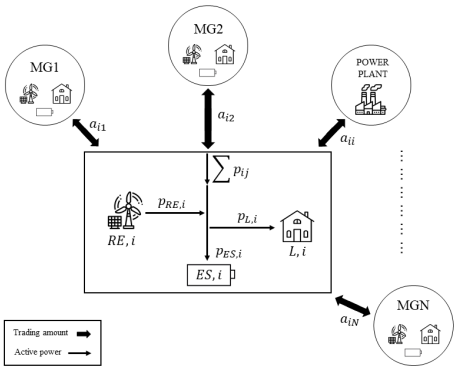

Fig. 1. P2P energy trading model for MGs.

world data sets to test the algorithm during four seasons.

II. System Model

We consider that there are N MGs and one power plant in the local area. Each MG has its renewable generators, ESSs and active loads. The MGs are connected with each other and the power plant by transmission lines operated by the distributed network operator. We assume the MG can make full use of its generators and storage system so that it can decide how to charge or discharge the battery and whether to turn down some of its generators if needed. Moreover, the MGs can observe the generation and demand meter and its battery level at each trading block. The MGs mainly depend on the renewable generation and storage system to meet their local energy demands; however, due to the intermittent nature of renewable energy, they need to trade their energy with each other or the power plant to balance the generation and demand as shown in Fig. 1.

A. Trading Strategy

We assume the energy trading takes place in the local hour-ahead P2P energy market, in which each trading block has one hour. At the beginning of each trading block, the MG will forecast its renewable generation and load demand based on historical data in the trading block. The amount of renewable energy of MG i in trading block t is denoted as $R_i(t)$, and the estimated generation is denoted as $\hat{R}_i(t)$. The actual and estimated amount of energy demand of MG i in trading block t are denoted as $D_i(t)$ and $\hat{D}_i(t)$, respectively. The remaining battery level of MG i at the beginning of trading block t is denoted as $S_i(t)$.

The strategy list of MG i is denoted as $\boldsymbol{x}_i(t) = [x_{ij}(t)]_{1\le j\le N, i\ne j} = [x_{i1}(t), x_{i2}(t), \dots x_{iN}(t)]$, where $x_{ij}(t)$ is the intended amount of energy trading from MG i to MG j in trading block t. If $x_{ij}(t) > 0$, which means MG i want to buy energy from MG j; if $x_{ij}(t) < 0$, which means MG i want to sell energy to MG j. Since MGs often have conflicting trading intentions, e.g., $x_{ij}(t) \times x_{ji}(t) > 0$, trading negotiations have been made, which resulting in actual trading action $\boldsymbol{a}_i(t) = [a_{ij}(t)]_{1\le j\ne i\le N} = [a_{i1}(t), a_{i2}(t), \dots a_{iN}(t)]$, where $a_{ij}(t) > 0$ means MG i buy energy from MG j; $a_{ij}(t) < 0$ means MG i sell energy to MG j. MGs only have a deal when one of them wants to sell energy and another wants to buy energy. It is clear that the actual energy trading might not be the same as the intention, therefore MGs need to buy or sell energy to the power plant to realize their strategy in trading block t. The amount of energy trading with the power plant in trading block t is denoted by $a_{ii}(t)$, which is the difference between the sum of $x_{ij}(t)$ and $a_{ij}(t), i \ne j$. Note that, the reason we denote it by $a_{ii}(t)$ is for algorithm convenience and we can use the vacant position $a_{ii}(t)$ to represent trading with the power plant and making just single list of $\boldsymbol{a}_i(t)$. The actual amount of energy trading of MG i is shown in (1).

$$a_{ij}(t) = \begin{cases} \frac{x_{ij}(t)}{|x_{ij}(t)|} \times \min\left(|x_{ij}(t)|, |x_{ji}(t)|\right), & \text{if } x_{ij} \times x_{ji} < 0, \forall i \ne j \\ 0, & \text{if } x_{ij} \times x_{ji} \ge 0, \forall i \ne j \\ \sum_{j=1}^{N} x_{ij}(t) - \sum_{j=1, j\ne i}^{N} a_{ij}(t), & \forall i = j. \end{cases} \tag{1}$$

B. Physical Constraints

For MG i, it will send or receive $a_{ij}(t)$ (kWh) energy in trading block t, which means that it will send or receive $p_{ij}(t) = a_{ij}(t)/T$ (kW) power in trading block t, where T is equal to 1 hour. In this model, we consider the transmission losses between MGs and other physical constraints. The transmission losses considered in this model are related to the electricity power, voltage and resistance. The resistance of a transmission line is proportional to the distance between MGs. Thus, when receiving power from other MGs, the real power received for MG i is $p_{ij}(t) - k_{ij} d {p_{ij}}^2(t)$, where k_{ij} is the loss constant, and d is the distance between MG i and MG j. The physical constraints can be written as

$$p_{ij}(t)^{\min} \le p_{ij}(t) \le p_{ij}(t)^{\max} \tag{2}$$

$$p_{ES,i}(t)^{\min} \le p_{ES,i}(t) \le p_{ES,i}(t)^{\max} \tag{3}$$

$$0 \le S_i(t+1) \le B \tag{4}$$

$$\sum_{j=1, \forall p_{ij}(t)>0}^{N} \left(p_{ij}(t) - k_{ij} d p_{ij}(t)^2\right) + \sum_{j=1, \forall p_{ij}(t)\le 0}^{N} p_{ij}(t) + p_{RE,i}(t) = p_{ES,i}(t) + p_{L,i}(t), \tag{5}$$

where $p_{RE,i}(t)$, $p_{ES,i}(t)$, $p_{L,i}(t)$, B are power from renewable generators, ESS, load device(kW) and capacity of the ESS (kWh) respectively.

The first three components are hard constraints, where (2) limits the power that MG i can receive from other MGs or power plant, (3) limits the power when charging or discharging the ESS battery, and (4) means that at the end of trading block t, the remaining ESS level cannot surpass its capacity. Constraint (5) means the MG must balance the energy generation

and consumption in trading block t. When charging the ESS, $p_{ES,i}(t) > 0$; when discharging the ESS, $p_{ES,i}(t) < 0$. In order to derive $S_i(t+1)$ in (4), the ESS is modeled as

$$S_i(t+1) = S_i(t) + E_{ch}\eta_{ch} - \frac{E_{dis}}{\eta_{dis}}, \tag{6}$$

where $E_{ch}(E_{dis})$, $\eta_{ch}(\eta_{dis})$ are the energy charging (discharging from) the battery and the charge (discharge) efficiency. Since charge and discharge action will degrade the condition of the batteries in the ESS, we consider the ESS wear cost, which will affect the energy trading strategies of the MGs. The empirical wear cost efficiency c_w (\$/kWh) [15] is shown as

$$c_w = \frac{C_{rep}}{S_b Q_b \sqrt{\eta_{rt}}}, \tag{7}$$

where C_{rep} is the replacement cost of the ESS, S_b is the battery size of the ESS, Q_b (kWh) is the lifetime of a battery unit in the storage and η_{rt} is the battery round-trip efficiency which is equal to the square of the storage discharge efficiency.

C. Utility Function

The utility function can help an MG evaluate the strategies that have been created in order to produce better strategies later. The reward or utility of MG i performing energy trading in trading block t, denoted as $u_i(t)$, depends on the trading profits, wear cost of the ESS, penalty if local demand is not met and virtual penalty if the MG wants to fulfill a certain goal. The local P2P market price can be dynamically changing, however, for encouraging MGs to trade energy with each other, the P2P energy trading price $\rho^-_{grid} \ll \rho^-_{p2p} \approx \rho^+_{p2p} \ll \rho^+_{grid} < \rho_{retail}$, where these symbols are the price MG selling energy to the power plant, other MGs, buying energy from other MGs, power plant and selling to the local consumers respectively. The utility function is expressed as

$$\begin{aligned} u_i(t) = & \sum_{j=1, j\neq i}^{N} a_{ij}(t)\left(I_{(a_{ij}\leq 0)}\rho^-_{p2p} - I_{(a_{ij}>0)}\rho^+_{p2p}\right) \\ & + a_{ii}(t)\left(I_{(a\leq 0)}\rho^-_{grid} - I_{(a_{ii}>0)}\rho^+_{grid}\right) + \rho_{\text{retail}} \times p_{L,i}(t)T \\ & - c_w\left|S_i(t+1) - S_i(t)\right| - C_{pen} - C_{vir}, \end{aligned} \tag{8}$$

where

$$C_{pen} = C_p p_{dif}(t) \tag{9}$$

$$C_{vir} = C_v a_{ii}(t). \tag{10}$$

The first term in the right-hand side of (8) is the trading profit of MG i trading with other MGs, the second term is the trading profit of MG i trading with the power plant, the third term is retail profit, the rest are energy storage wear cost and other penalties. The demand penalty C_{pen} happens when $\sum p_{ij} + p_{RE,i} < p^{min}_{ES,i} + p_{L,i}$, where $p_{dif} = p^{min}_{ES,i} + p_{L,i} - \sum p_{ij} + p_{RE,i}$ and C_p is the penalty coefficient. To be noticed that if $\sum p_{ij} + p_{RE,i} > p^{max}_{ES,i} + p_{L,i}$, MG i can always reduce their generation output or selling to the grid to balance the demand. The virtual penalty C_{vir} is optional, and its existence is to make the algorithm believe achieving some goal is beneficial even though it might not be economically optimal. In this paper, the objective of MG i is to maximize the trading profits while also minimizing the dependence on the power plant. Thus, the virtual penalty can be set as (10), where C_v is a virtual coefficient. The virtual penalty can be also set to achieve other social welfare goals for the MG.

D. System Problem

As each MG does not know energy generation and demand information of other MGs, MG i will choose its trading strategies $\boldsymbol{x}_i(t)$ based on the estimated generation $\hat{R}_i(t)$, energy demand $\hat{D}_i(t)$ and current storage level $S_i(t)$. Therefore, the utility function can also be written as

$$u_i(t) = u\left(\hat{R}_i(t), \hat{D}_i(t), S_i(t) | \boldsymbol{x}_i(t)\right). \tag{11}$$

The goal is to maximise the expected total utility which is the sum of all future utilities based on the optimal policy $\pi\left(\boldsymbol{x}_i(t) | \hat{R}_i(t), \hat{D}_i(t), S_i(t)\right)$, which can be shown as

$$\mathbf{P1}: \max_{\pi} U_{i_\pi}(t) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau u_i(t+\tau+1)\right]. \tag{12}$$

The trading policies made by MGs could be based on naive intention (Trading surplus or needed energy of trading block t without thinking about the future), board resolution, or an automatic energy management system (AEMS). In this paper, MG i will use deep Q-learning algorithm as part of an AEMS to derive better strategies over time.

III. Deep Reinforcement Learning and Solution Algorithm

DRL is the combination of deep learning (DL) and reinforcement learning (RL), where RL is a mathematical framework for experience-driven behaviour learning, and DL consists of deep neural networks which can be used as function approximators in RL. P2P energy trading involves a large number of continuous data sets in which are made up of stochastic and uncertain data like renewable generation and load demand, so making a decision by human or conventional optimisation methods would be challenging. With DRL, making optimal decisions in P2P energy trading could be possible.

A. Deep Q-learning

Deep Q-learning, also called DQN algorithm, consists of deep neural networks (DNN) and Q-learning. The idea of deep Q-learning is to approximate the Q-values using DNN since the basic Q-learning cannot tackle the problems with high-dimensional state-action space and continuous data sets.

In the DQN updating function (13), the target value function $Q(s,a)$ is replaced by a parameterized value function [16] $Q(s,a;\theta)$, where θ is the parameters that define the Q-values, $\max_a Q(s',a';\theta')$ is the estimate of optimal future value.

$$Q^{new}(s,a;\theta) \leftarrow Q(s,a;\theta) + \alpha\left(R(s,a) + \gamma \max_{a} Q\left(s',a';\theta'\right) - Q(s,a;\theta)\right). \quad (13)$$

The action is chosen following an ε-greedy policy, while the updates of parameters are made on a randomly selected mini-batch which is a set of transitions (s,a,r,s'), results in less variance than just updating a single tuple. This experience replay technique allows the algorithm to explore a large range of previous state-action space; otherwise, DNN tends to rewrite them with new experiences. The updates equation of parameters and details of experience replay will be shown in the energy trading algorithm section.

B. DQN-based P2P Energy Trading algorithm

First, we need to input the state into DNN at the beginning of the trading block t. The observed state before trading block t is $\left[\hat{R}_i(t), \hat{D}_i(t), S_i(t)\right]$. As the state in trading block t is not fully observable, we formulate an experience sequence $\boldsymbol{\varphi}(t)$ consisting of the current estimated state and last fully observed state-action pair, with $\boldsymbol{\varphi}_i(t) = \left(R_i^-, D_i^-, \boldsymbol{a}_i^-, \hat{R}_i, \hat{D}_i, S_i\right)$. The input of the DNN with parameters in trading block t is denoted by $\boldsymbol{\theta}_t$, the output of the DNN is $Q\left(\boldsymbol{\varphi}_i(t), \boldsymbol{x}_i(t); \boldsymbol{\theta}_t\right)$, and the trading strategy for MG i is chosen based on ε-greedy policy. With probability ε, the strategy is selected randomly, otherwise selecting the strategy that maximizes the Q-value.

After evaluating the trading utility in trading block t and getting the new experience sequence $\boldsymbol{\varphi}_i(t+1)$, the algorithm stores the transition $\left(\boldsymbol{\varphi}_i(t), \boldsymbol{x}_i(t), u_i(t), \boldsymbol{\varphi}_i(t+1)\right)$ in the replay memory pool $\mathbb{D}$. The next step is to sample random transition from $\mathbb{D}$, and the parameters $\boldsymbol{\theta}_t$ are updated byminimizing the loss function shown in (14) using gradient descent. Note that, the parameters $\boldsymbol{\theta}_t^-$ remain the same as $\boldsymbol{\theta}_t$ and are only update every C iterations to reduce the risk of divergence. The pseudocode of P2P energy trading for MGs is shown in Algorithm 1.

$$L(\boldsymbol{\theta}_t) = \mathbb{E}_{\left(\boldsymbol{\varphi}_i, \boldsymbol{x}_i, u_i, \boldsymbol{\varphi}_i'\right) \sim U(\mathbb{D})} \left[\left(u_i + \gamma \max_{\boldsymbol{x}_{i'}} Q\left(\boldsymbol{\varphi}_i', \boldsymbol{x}_i'; \boldsymbol{\theta}_t^-\right) - Q\left(\boldsymbol{\varphi}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_t\right)\right)^2\right]. \quad (14)$$

IV. Simulation Results

In this section, the deep Q-learning for P2P energy trading algorithm was simulated by using real data from Pecan Street Inc. [17], which consists of 1-year electricity generation and demand data at 1-hour resolution from 100 households located in Mueller, Austin, Texas. The 100 households were divided into three groups as three MGs, the PV generation of the households was aggregated properly to work as a sufficient renewable generator for local MG. Also, the P2P electricity prices followed hourly LMPs records from ISO New England Inc. [18]. The system parameters are given in Table I. with ESS parameters given in Table II.

Algorithm 1: Deep Q-Learning for P2P Energy Trading

Initialize γ, θ_1 and replay memory $\mathbb{D}$ to capacity N_{max}
for $t \in \mathbf{T}$ **do**
  Forecast $\hat{R}_i(t), \hat{D}_i(t)$ and observe $S_i(t)$
  Form experience sequence $\boldsymbol{\varphi}_i(t)$
  Input $\boldsymbol{\varphi}_i(t)$ with $\boldsymbol{\theta}_t$ and get $Q\left(\boldsymbol{\varphi}_i(t), \boldsymbol{x}_i(t); \boldsymbol{\theta}_t\right)$
  Choose trading strategy $\boldsymbol{x}_i(t)$ using ε-greedy
  for $j \in \mathbb{N}$ **do**
    Receive the intended energy $x_{ji}(t)$ from MG j
  end
  Calculate $a_{ij}(t)$ via (1) and Check constrain $p_{ij}(t)$
  Observe actual generation $R_i(t)$ and demand $D_i(t)$
  Calculate constrain $p_{ES,i}(t)$ via (5)
  if $p_{ES,i}(t)$ *not in constraint (3)* **then**
    $p_{ES,i}(t) = p_{ES,i}^{limit}(t)$
  end
  Calculate $S_i(t+1)$ via (6)
  if $S_i(t+1)$ *not in constraint (4)* **then**
    $S_i(t+1) = S_i^{limit}(t)$
  end
  Calculate Penalty using (9), (10)
  Observe the electricity price $\rho_{grid}, \rho_{p2p}, \rho_{retail}$
  Calculate utility $u_i(t)$ via (8)
  Store transition $\left(\boldsymbol{\varphi}_i(t), \boldsymbol{x}_i(t), u_i(t), \boldsymbol{\varphi}_i(t+1)\right)$ in $\mathbb{D}$
  Calculate loss function $L(\boldsymbol{\theta}_t)$ via (14)
  Update DNN parameters $\boldsymbol{\theta}_t$ by gradient descent
end

TABLE I
System Parameters

Parameters	Values	
Power limit (kW)	$-150 \le p_{12} \le 150$	$-200 \le p_{13} \le 200$
Distance (km)	$d_{12} = 50$	$d_{13} = 100$
Loss constant	$k_{12} = 6.66 \times 10^{-6}$	$k_{13} = 5 \times 10^{-6}$
Penalty coefficient	$C_p = 0.3$	$C_v = 0.2$
Electricity price	$\rho_{\text{grid}}^- = 0.8\rho_{P2P}, \rho_{grid}^+ = 1.2\rho_{P2P}, \rho_{retail} = 1.8\rho_{P2P}$	

We choose MG1 as our agent, which consists of 30 households. The one-year PV generation and local demand for MG 1 are shown in Figure 2. The DQN-based trading strategy with other MGs in each hour is from -150 kWh to 150 kWh at 30 kWh step. Thus, the number of total strategies with other 2 MGs is 121. As the action space is impossible for a basic Q-learning, we design a rule-based trading strategy as a benchmark. The rule-based trading strategy is to sell estimated surplus energy or buy estimated needed energy in the next

TABLE II
Battery Parameters

Battery Model	A	B	C
Capacity	300 kWh	400 kWh	500 kWh
Rated Power	80 kW	100 kW	130 kW
Wear Cost	0.009$/kWh		
Efficiency	$\eta_{ch} = \eta_{dis} = 0.9$		

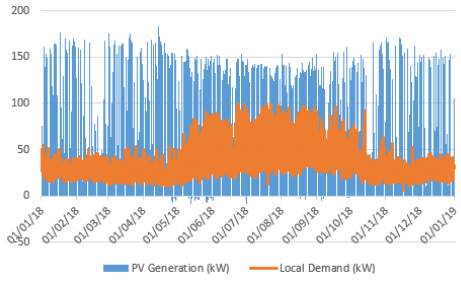

Fig. 2. PV generation and local demand for MG 1.

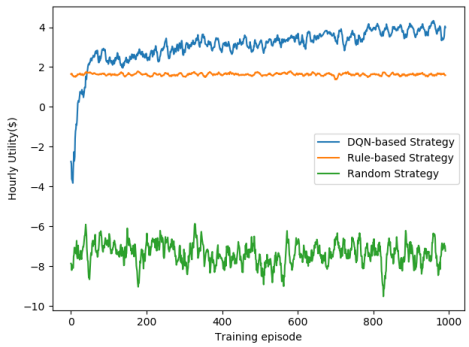

Fig. 3. The hourly utility of MG 1 in P2P energy trading.

trading time. There is also a random strategy that the MG will choose the trading action randomly. Figure 3. shows that the DQN-based strategy outperforms other strategies.

ESS is essential in P2P energy trading. Therefore the impact of different battery sizes (shown in Table II.) during the four seasons is studied and shown in Figure 4. With no battery, the utility of MG 1 is 37 per cent lower than having the battery A. However, the result shows that larger battery size is not always better. The larger-size battery may result in a massive amount of charge and discharge and resource waste as we consider charge and discharge rate and battery wear cost. The utilities of MG1 are about the same in spring and winter; while in summer and autumn, the utilities drop. This is because MGs are busy meeting their own demand. In addition, we found adding the virtual penalty C_{vir} can reduce the power plant schedule by 82 per cent although it is not economically beneficial. To conclude, the proposed DQN-based energy-trading model can choose better trading strategies to improve the utility across seasonal changes. Meanwhile, it can also help MGs to choose the most suitable battery and achieve their own social goals.

V. Conclusion

In this paper, we proposed a P2P energy trading for MGs using DRL. With several essential physical constraints, the model can be better adapted for real situations. The simulation was performed using 1-year real generation and demand data, showing that the proposed DQN-based energy-trading model can choose better trading strategies to improve the utility across seasonal changes. This model can also help MGs to choose the most suitable battery and achieve their own social goals.

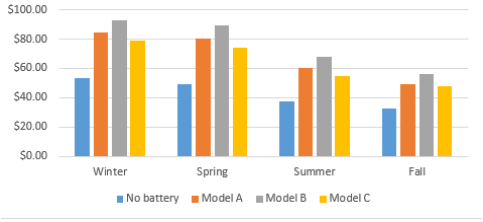

Fig. 4. Average daily utilities with different battery models.

References

[1] W. Tushar, C. Yuen, H. Mohsenian-Rad, T. Saha, H. V. Poor, and K. L. Wood, "Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 90–111, mar 2018.

[2] Y. Wang, W. Saad, Z. Han, H. V. Poor, and T. Baar, "A game-theoretic approach to energy trading in the smart grid," *IEEE Transactions on Smart Grid*, vol. 5, no. 3, pp. 1439–1450, may 2014.

[3] C. Long, J. Wu, C. Zhang, L. Thomas, M. Cheng, and N. Jenkins, "Peer-to-peer energy trading in a community microgrid," in *IEEE Power and Energy Society General Meeting*, vol. 2018-Janua, 2018, pp. 1–5.

[4] S. Park, J. Lee, G. Hwang, and J. K. Choi, "Event-Driven Energy Trading System in Microgrids: Aperiodic Market Model Analysis with a Game Theoretic Approach," *IEEE Access*, vol. 5, pp. 26 291–26 302, 2017.

[5] G. El Rahi, S. R. Etesami, W. Saad, N. B. Mandayam, and H. V. Poor, "Managing Price Uncertainty in Prosumer-Centric Energy Trading: A Prospect-Theoretic Stackelberg Game Approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 702–713, 2019.

[6] W. Kou and S. Y. Park, "Game-theoretic approach for smartgrid energy trading with microgrids during restoration," in *IEEE Power and Energy Society General Meeting*, vol. 2018-Janua. IEEE, jul 2018, pp. 1–5.

[7] T. Morstyn, A. Teytelboym, and M. D. McCulloch, "Bilateral contract networks for peer-to-peer energy trading," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2026–2035, 2019.

[8] R. Li, W. Wei, S. Mei, Q. Hu, and Q. Wu, "Participation of an Energy Hub in Electricity and Heat Distribution Markets: An MPEC Approach," *IEEE Transactions on Smart Grid*, vol. 3053, no. c, pp. 1–13, 2018.

[9] V. François-lavet, R. Fonteneau, and D. Ernst, "Deep Reinforcement Learning Solutions for Energy Microgrids Management," in *European Workshop on Reinforcement Learning*, 2016, pp. 1–7.

[10] G. K. Venayagamoorthy, R. K. Sharma, P. K. Gautam, and A. Ahmadi, "Dynamic Energy Management System for a Smart Microgrid," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 8, pp. 1643–1656, aug 2016.

[11] B. V. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck, "Battery energy management in a microgrid using batch reinforcement learning," *Energies*, vol. 10, no. 11, pp. 1–19, 2017.

[12] Y. Chen, W. Wei, F. Liu, E. E. Sauma, and S. Mei, "Energy Trading and Market Equilibrium in Integrated Heat-Power Distribution Systems," *IEEE Transactions on Smart Grid*, vol. PP, no. c, p. 1, 2018.

[13] L. Xiao, X. Xiao, C. Dai, M. Pengy, L. Wang, and H. V. Poor, "Reinforcement Learning-based Energy Trading for Microgrids," 2018.

[14] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," sep 2015.

[15] S. Han, S. Han, and H. Aki, "A practical battery wear model for electric vehicle charging applications," *Applied Energy*, vol. 113, pp. 1100–1108, 2014.

[16] V. Francois-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, *An Introduction to Deep Reinforcement Learning*, 2018.

[17] "Source: Pecan Street Inc. Dataport 2018." [Online]. Available: https://www.pecanstreet.org/

[18] "Electricity price." [Online]. Available: https://www.iso-ne.com/