Deep Reinforcement Learning based Charging Pricing for Autonomous Mobility-on-Demand System

Lu, Ying; Liang, Yanchang; Ding, Zhaohao; Wu, Qiuwei; Ding, Tao; Lee, Wei Jen

Published in:
IEEE Transactions on Smart Grid

Link to article, DOI:
10.1109/TSG.2021.3131804

Publication date:
2022

Document Version
Peer reviewed version

Deep Reinforcement Learning based Charging Pricing for Autonomous Mobility-on-Demand System

Ying Lu, *Student Member, IEEE,* Yanchang Liang, *Student Member, IEEE,* Zhaohao Ding*, *Senior Member, IEEE,* Qiuwei Wu, *Senior Member, IEEE,* Tao Ding, *Senior Member, IEEE,* Wei-Jen Lee, *Fellow, IEEE*

***Abstract*—The autonomous mobility-on-demand (AMoD) system plays an important role in the urban transportation system. The charging behavior of AMoD fleet becomes a critical link between charging system and transportation system. In this paper, we investigate a strategic charging pricing scheme for charging station operators (CSOs) based on a non-cooperative Stackelberg game framework. The Stackelberg equilibrium investigates the pricing competition among multiple CSOs, and explores the nexus between the CSOs and AMoD operator. In the proposed framework, the responsive behavior of AMoD operator (order-serving, repositioning, and charging) is formulated as a multi-commodity network flow model to solve an energy-aware traffic flow problem. Meanwhile, a soft actor-critic based multi-agent deep reinforcement learning algorithm is developed to solve the proposed equilibrium framework while considering privacy-conservation constraints among CSOs. A numerical case study with city-scale real-world data is used to validate the effectiveness of the proposed framework.**

***Index Terms*—EV charging pricing, deep reinforcement learning, power and transportation system, autonomous mobility-on-demand, soft actor-critic.**

NOMENCLATURE

A. Sets, Index, and Tuples

D	Set of days indexed by d.
T	Set of horus indexed by t.
C	Set of EV state-of-charge (SoC) indexed by c.
$\mathcal{V}_R$	Set of transportation nodes indexed by i.
$\mathcal{V}_g$	Set of transportation nodes in an augmented graph indexed by I, J or Q.
$\mathcal{E}_R$	Set of transportation roads indexed by $(I,J) \in \mathcal{E}_R : i_I \neq i_J$ or $(Q,I) \in \mathcal{E}_R : i_Q \neq i_I$.
$\mathcal{E}_C$	Set of charging process at the station in an augmented graph indexed by $(I,J) \in \mathcal{E}_R : i_I = i_J = n$ or $(Q,I) \in \mathcal{E}_R : i_Q = i_I = n$.
$\mathcal{E}_g$	Set of transportation roads in an augmented graph, which is partitioned into two subsets, namely $\mathcal{E}_g = \mathcal{E}_R + \mathcal{E}_C$.
K	Set of transportation requests indexed by k.
M	Set of charging station operators (CSOs) indexed by m.
N	Set of electric vehicle charging stations (EVCSs) indexed by n.
I	Tuple for transportation nodes $I = (i_I, t_I, c_I) \in \mathcal{V}_g$ in an augmented graph. i_I is a node in the road network. t_I is a discrete-time. c_I is a discrete SoC.
k	Tuple for transportation requests $k = (o_k, d_k, t_k, \lambda_k)$, where $o_k, d_k \in \mathcal{V}_R$ represent the request's origin and destination location, respectively. t_k is the requested pickup time, and λ_k is the number of requests k.

B. Parameters

C_n^c	The vehicle capacity of EVCS n [vehicles].
$C_{(i,j)}^r$	The vehicle capacity of transportation road (i_I, i_J) [vehicles].
V^T	Value of time [$/hour].
V^D	Value of distance [$/km].
P^C	Energy equivalent to one SoC [kW/p.u.].
$T_{(i,j)}$	Traveling time of road (i_I, i_J) [hour].
$D_{(i,j)}$	Distance of road (i_I, i_J) [km].
$\alpha_{n,t}$	Distribution Locational marginal prices (DLMPs) of EVCS n at time t [$/MWh].
η	The chargers' efficiency.

C. Variables

$x_{k,(I,J)}^{Ser}$	The order-serving flow belonging to request k traveling from location i_I to location i_J from time t_I to time t_J, with an initial SoC of c_I and SoC of c_J [vehicles].
$x_{(I,J)}^{Rep}$	The repositioning flow represents the number of empty EV traveling from location i_I to location i_J or charging at location $i_I = i_J$ from time t_I to time t_J, with an initial SoC of c_I and a final SoC of c_J [vehicles].
$x_{(I,J)}^{Cha}$	The charging flow represents the number of empty EV traveling from location i_I to location i_J or

This work is supported in part by the National Key RD Program of China under Grant 2019YFE0118400, in part by the National Natural Science Foundation of China under Grant 51907063, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2020QNRC001, and in part by the China Scholarship Council under Grant 202006730008. *(Corresponding author: Zhaohao Ding.)*

Ying Lu, Yanchang Liang and Zhaohao Ding are with the School of Electrical and Electronic Engineering, North China Electric Power University, Beijing 102206, China (e-mail: yinglu@elektro.dtu.dk; yanchang.liang@warwick.ac.uk; zhaohao.ding@ncepu.edu.cn).

Qiuwei Wu is with the Center for Electric Power and Energy, Department of Electrical Engineering, Technical University of Denmark, 2800 Lyngby, Denmark (e-mail: qw@elektro.dtu.dk).

Tao Ding is with the Department of Electrical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China (e-mail: tding15@xjtu.edu.cn).

Wei-Jen Lee is with the Energy Systems Research Center, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: wlee@uta.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3131804, IEEE Transactions on Smart Grid

charging at location $i_I = i_J$ from time t_I to time t_J, with an initial SoC of c_I and a final SoC of c_J [vehicles].

$z^{\text{ori}}_{k,c}$ The order-serving flow departing from the origin location o_k with SoC c for request k at time t_k [vehicles].

$z^{des}_{k,c,t}$ The order-serving flow reaching the destination location d_k with SoC c for request k at time t [vehicles].

$\pi^{CS}_{n,t}$ Charging price of EVCS $n \in N$ at time t [$/MWh].

$\pi^{CSO}_{n_m,t}$ Charging price of EVCS $n_m \in N_m$ managed by CSO m at time t [$/MWh].

$p^{EV}_{n,t}$ Charging loads of EVCS n at time t [kW].

$p^{EV}_{n_m,t}$ Charging loads of EVCS $n_m \in N_m$ managed by CSO m at time t [kW].

$p^{grid}_{n,t}$ Energy procurement of EVCS n at time t [kW].

κ Dual variables of AMoD problem.

I. Introduction

WITH the move of transportation electrification [1] and sharing economy [2], the autonomous mobility-on-demand (AMoD) fleet has become one of the most transformative and promising transportation modes [3]. Through its autonomous and shared characteristics, the deployment of electric vehicles (EVs) based AMoD system can increase overall vehicle utilization (one AMoD vehicle can replace seven private-owned vehicles [4]), minimize public parking demand, and reduce environmental pollution [5]. Furthermore, with multiple companies now heavily investing in AMoD technology, it could become one of the dominant urban transportation modes in the near future [6].

Generally speaking, the AMoD fleet could coordinate its routing and charging schedules more efficiently than conventional private electric vehicles (EVs) as it is managed by a centralized operator. Typically, AMoD operator needs to make three types of decisions [5], which are order-serving (fulfilling traveling requests with the specific origin and destination nodes and starting time), vehicle reposition (repositioning vehicles to a certain region in advance for future needs), and charging (determining the charging location and time for each vehicle). Therefore, the behavior pattern of the AMoD fleet is more predictable as it is centralized determined by the operator, which maximizes the total operating profit. In contrast with that, private EV users rely on individually rational. They would determine their behaviors based on personal preference, which contains much higher randomness than a centralized profit-driven AMoD operator. Considering those unique characteristics, the optimal operation of AMoD fleet has been explored by multiple researchers, focusing on picking up passengers, routing to destinations, and repositioning idle vehicles for future order-serving or charging decisions. For example, Pavone *et al.* [7] developed fluidic-based methods to ensure AMoD operators make the optimal fleet management decisions, and they further presented queueing network methods for maximizing the throughput of an AMoD urban transportation system in [8]. Rossi *et al.* [9] modeled the AMoD fleet routing problem within a network flow framework without increasing traffic congestion. Turan *et al.* [10] defined the dynamic system model that captures the time-dependent and stochastic features of the AMoD system. Iglesias *et al.* [11] present a data-driven framework to control AMoD fleets where the Model Predictive Control algorithm is adopted to leverage demand forecasts. Cocca *et al.* [12] developed a discrete-event trace-driven simulator to study the usage of an EV sharing system. Guériau *et al.* [13] proposed a reinforcement learning-based decentralized approach to vehicle relocation as well as ride request assignment in shared AMoD systems, and Swaszek *et al.* [14] proposed a parametric threshold-based control driven by the known relative abundance of AMoD vehicles. Those literatures above explore the optimal order serving and repositioning decisions for AMoD operator while the charging scheduling problem is barely considered.

With the potential large-scale implementation in the near future, the charging behavior of the AMoD fleet could become a critical link between power systems and transportation systems. Consequently, strategic charging demand management techniques, such as charging service pricing, can not only affect the operation boundary conditions of transportation system via altering the scheduling decision of AMoD operator, but also change the operation efficiency of power systems by reshaping the charging load distribution in both spatial and temporal manner. The price-based charging management technique of CSOs has been investigated by many researchers. For example, He *et al.* [15] applied a Lagrangian relaxation-based iterative scheme to design congestion tolls and locational marginal prices for the power-transportation system. The author further proposed a multi-class combined distribution and assignment model to optimize electricity prices and road tolls, minimizing power losses and traveling times in [16], where the power system is described by the alternating current optimal power flow. Wei *et al.* [17] proposed network equilibrium of the coupled power-transportation system to analyze the optimal traffic-power flow and calculate the locational marginal prices. Although those works significantly contribute to the price-based management for charging demand, most of them do not consider the unique operational characteristics of the AMoD system as its order serving, vehicle repositioning, and charging are centralized coordinated.

Furthermore, it is also essential to consider the behavior patterns of competitive CSOs when considering the pricing-based charging demand management as the responsive behavior of AMoD is also affected by the charging service price of other CSOs. Numerous studies investigated the price competition of CSOs. Lee *et al.* [18] investigated the price competition among CSOs by using a Stackelberg game. Yuan *et al.* [19] considered the competitive pricing problem of each CSO based on the prediction of CSO selection decisions and the other station's pricing decisions. Ghosh *et al.* [20] and Moradipari *et al.* [21] developed menu-based pricing schemes for EV charging access to multiple EVCSs.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3131804, IEEE Transactions on Smart Grid

Moghaddm *et al.* [22] presented a pricing model among CSOs to reduce the overlaps between residential and charging loads by drifting EVs to less popular stations. Zhang *et al.* [23] proposed a pricing model of battery switching stations by establishing the detailed EV and battery agents through coding their states and transitions. Those works explore the competition characteristics among CSO, but they hardly consider the spatial-temporal responsive behavior pattern of AMoD in their models.

In this paper, we propose a strategic charging pricing scheme for charging station operators (CSOs) based on a non-cooperative Stackelberg equilibrium framework. The equilibrium framework is established to investigate the pricing competition among CSOs and the nexus between charging infrastructure and AMoD system. An equilibrium problem with equilibrium constraints (EPEC) is formulated to model the proposed framework. Due to the computational complexity introduced by the inherent non-convexities and non-linearity and complete information constraints introduced by the privacy concern of each CSO, it presents challenges for conventional methods, such as [18], [19], [24], to solve the proposed EPEC problem. To overcome those challenges, a multi-agent deep reinforcement learning (MADRL) framework based on a soft actor-critic (SAC) algorithm is constructed. In the DRL solution framework, CSOs (agents) learn their optimal pricing strategy in an interactive environment by trial and error using feedback from their actions and experiences [25]. That is, as long as the objective function can be calculated given a set of observations, the charging pricing strategy of CSOs (agents) can be designed and adjusted. Through trial-and-error interactions within a dynamic environment, such learning-based approaches for CSOs avoid the significant modeling and computational complexity posed by EPEC models. Furthermore, no internal information (such as computational algorithm and operating parameters) from their competitors and the AMoD system is required by the strategic CSO. Hence, the privacy of each CSO is preserved. During competitive equilibrium, CSOs only rely on their operating parameters, the trip requests, and the non-proprietary information of distribution locational marginal prices (DLMPs). In the proposed framework, the responsive behavior of AMoD operator (coordinated order-serving, repositioning, and charging decisions) is formulated as a multi-commodity network flow model to solve an energy-aware traffic flow problem. The proposed model is designed to make short-term decisions with a fixed setting on EVCS location and capacity as the siting and sizing problem is out of the scope of this paper. The major contributions of this paper are summarized as follows:

(1) A strategic charging pricing scheme based on a non-cooperative Stackelberg equilibrium framework is proposed to support the charging service pricing of CSO towards AMoD system. The proposed equilibrium framework investigates the pricing competition among CSOs and the nexus between CSO decisions and AMoD decisions.

(2) A MADRL framework based on a SAC algorithm is developed to solve the proposed equilibrium framework with privacy-conservation constraints among CSOs. Meanwhile, a multi-commodity network flow model is formulated to characterize the unique responsive behavior of AMoD system.

(3) Numerical case studies with city-scale real-world data is provided to demonstrate the effectiveness and computational efficiency of the proposed framework.

The rest of the paper is organized as follows. Section II presents the non-cooperative Stackelberg equilibrium framework, and Section III proposes a deep reinforcement learning-based algorithm to solve it. Section IV presents experimental simulation results to illustrate the proposed method, followed by the conclusions provided in Section V.

II. Equilibrium Framework and Mathematical Formulation

As illustrated in Fig. 1, a strategic charging pricing scheme for CSOs based on a non-cooperative Stackelberg equilibrium framework is proposed. Stackelberg equilibrium are considered in this framework. The Stackelberg equilibrium investigates the interactive characteristics between charging network and AMoD system, while investigates the Nash equilibrium among competitive CSOs. Under this setting, profit-driven CSOs determine their charging pricing strategy for each charging station based on the power procurement cost and estimated charging demand. Meanwhile, the AMoD operator optimally manages its fleet by coordinate the order service, vehicle reposition, and charging schedule simultaneously, which effectively shapes the spatial and temporal charging demand distribution for CSOs. Therefore, the EV charging profiles of charging stations are flexible in time and space, which is determined by the charging schedule of AMoD operator. Combining those two equilibriums, it shall be noted that the charging price is playing a critical role in coordinating the charging network and AMoD system. In other words, charging prices determined by CSOs affect the AMoD operator's decision on when and where to charge their vehicles. Meanwhile, in response to those pricing decisions, the spatial-temporal distribution of charging demand can be altered, which affects the market share and profit of each CSO.

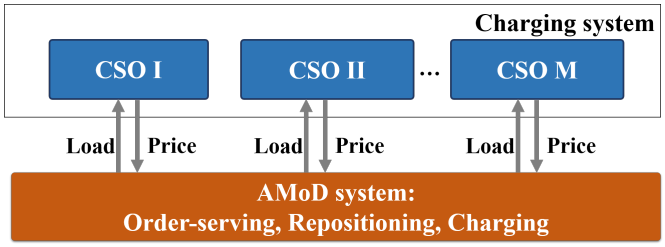


Fig. 1. Equilibrium framework for coupled charging system and AMoD system.

A. AMoD Fleet Operator Model

To characterize the interactive characteristics between AMoD system and charging network, it is essential to model the price responsive behavior of AMoD operator. As discussed earlier, multiple types of methods can be used to model the decision-making process of AMoD, from microscopic [26], [27] to macroscopic approaches [9], [28]. Inspired by [29], we adopt a multi-commodity network flow model to capture the price responsive behavior of AMoD system. In this model, the AMoD operator is assumed to determine the order serving routes (i.e., order-serving), reposition idle vehicles for future order-serving and charging needs (i.e., vehicle repositioning), and schedule charging time and location for vehicles (i.e., charging scheduling) in a given transportation network, which is modeled in the following.

1) Transportation Network

To characterize the operation of AMoD, we model the road network as a directed graph $\mathcal{G} = (\mathcal{V}_R, \mathcal{E}_R)$. The edges $(i,j) \in \mathcal{E}_R$ designated to represent major roads connecting nodes i and j where the nodes $i, j \in \mathcal{V}_R$ represent entrances or exits of a road, such as charging stations or trip terminals. To track the time and state of charge (SoC) dynamics of AMoD fleet, we expand the road network graph to an augmented energy-time-space graph $\alpha\mathcal{G} = (\mathcal{V}_g, \mathcal{E}_g)$. Due that the node set $\mathcal{V}_g$ is the extension of $\mathcal{V}_R$, a node $I \in \mathcal{V}_g$ corresponds to a tuple $I = (i_I, t_l, c_I)$, where $i_I \in \mathcal{V}_R$ is a node in the road network graph $\mathcal{G}$; $t_I \in \{1, \cdots, T\}$ is a discrete-time; and $c_I \in \{1, \cdots, C\}$ is a discrete charge level. The edge set $\mathcal{E}_g$ is partitioned into two subsets, namely $\mathcal{E}_g = \mathcal{E}_C + \mathcal{E}_R$. Edges $(I,J) \in \mathcal{E}_R$ represent road links, whereas edges $(I,J) \in \mathcal{E}_C$ model the charging process at the stations. To facilitate such formulation, the time and SoC in the proposed model are considered as discrete values. It should be mentioned that this assumption does not significantly affect the results considering the number of vehicles and the time scale of CSO pricing.

The transportation request k is represented by a set of tuple $k = (o_k, d_k, t_k, \lambda_k)$, where $o_k \in \mathcal{V}_R$ is the request origin node, $d_k \in \mathcal{V}_R$ is the request destination node, $t_k \in T$ is the request starting time, and $\lambda_k \in \mathbb{R}$ is the travel request. Formally, for the request k, the customer flow is a function $x^{Ser}_{k,(I,J)}$, which represents the rate of order-serving flows belonging to the request k traveling from location i_I to location i_J during the time t_I and t_J, with an initial SoC c_I and a final SoC c_J. Analogously, the repositioning flow $x^{Rep}_{(I,J)}\,(\forall (I,J) \in \mathcal{E}_C, i_I \neq i_J)$ represents the rate of idle vehicles traversing a road. $x^{Cha}_{(I,J)}\,(\forall (I,J) \in \mathcal{E}_R, i_I = i_J = n)$ represents the charging flows from time t_I with an initial SoC c_I at the virtual charging node $n \in N$. Moreover, $z^{\text{ori}}_{k,c}$ represents start-service flows with a start SoC c belonging to the request k. $z^{\text{des}}_{k,c,t}$ represents end-service flows with a start SoC c at tine t belonging to the request k.

2) Mathematical Formulation

Based on the augmented energy-time-space graph, the fleet navigation problem for AMoD operator can be formulated as a multi-commodity network flow model, as shown in the following.

$$\min_{\Xi_1} \begin{bmatrix} \sum_{(I,J)\in\mathcal{E}_R,k} \left[x^{Ser}_{k,(I,J)} \left(V^T T_{(i,j)} + V^D D_{(i,j)}\right)\right] \\ + \sum_{(I,J)\in\mathcal{E}_R} \left[x^{Rep}_{(I,J)} \left(V^T T_{(i,j)} + V^D D_{(i,j)}\right)\right] \\ + \sum_{(I,J)\in\mathcal{E}_C} \left[x^{Cha}_{(I,J)} \left(c_J - c_I\right) P^C \pi^{CS}_{n,t}\right] \end{bmatrix}, \tag{1}$$

$$\Xi_1 = \left\{x^{Ser}_{k,(I,J)}, x^{Rep}_{(I,J)}, x^{Cha}_{(I,J)}, z^{des}_{k,c,t}, z^{ori}_{k,c}\right\}$$

$$\text{s.t.} \sum_J x^{Ser}_{k,(I,J)} + z^{des}_{k,c,t} 1_{i_I = d_k} = \sum_Q x^{Ser}_{k,(Q,I)} + z^{ori}_{k,c} 1_{i_I = o_k} 1_{t_l = t_k} \quad (\forall k, I) : \kappa^{SF}_{k,I}, \tag{2}$$

$$\sum_J x^{Rep}_{(I,J)} + \sum_J x^{Cha}_{(I,J)} + \sum_k z^{ori}_{k,c_I} 1_{i_I = o_k} 1_{t_I = t_k} = \sum_k z^{des}_{k,c_I,t_I} 1_{i_I = d_k} + \sum_Q x^{Rep}_{(Q,I)} + \sum_Q x^{Cha}_{(Q,I)} \quad (\forall I) : \kappa^{RF}_I, \tag{3}$$

$$\sum_c z^{ori}_{k,c} = \sum_c \sum_t z^{des}_{k,c,t} = \lambda_k \quad (\forall k) : \kappa^{SO}_k, \kappa^{SD}_k, \tag{4}$$

$$\sum_c x^{Ser}_{k,(I,J)} 1_{c_I < c_{ij}} = x^{Rep}_{(I,J)} 1_{c_I < c_{ij}} = 0 \quad (\forall (i_I, i_J) \in \mathcal{E}_R, t_I) : \kappa^{SS}_{(i_I,i_J)\in\mathcal{E}_R,t_I} \kappa^{RS}_{(i_I,i_J)\in\mathcal{E}_R,t_I}, \tag{5}$$

$$\sum_{(i_1,i_j)\in\mathcal{E}_c} x^{Cha}_{(I,J)} \leq C^{\text{c}}_n \quad (\forall n, t_I) : \kappa^{CR}_n, \tag{6}$$

$$\sum_{c_I,k} x^{Ser}_{k,(I,J)} + \sum_{c_I} x^{Rep}_{(I,J)} \leq \text{C}^r_{i,j} \quad (\forall (i_I, i_J) \in \mathcal{E}_R, t_I) : \kappa^{RR}_{(i_1,i_J)\in\mathcal{E}_R,t_I}. \tag{7}$$

The cost function (1) consists of three parts corresponding to three types of flows in the AMoD system: order-serving, repositioning, and charging. The first item $\sum_{(I,J)\in\mathcal{E}_R,k} \left[x^{Ser}_{k,(I,J)} \left(V^T T_{(i,j)} + V^D D_{(i,j)}\right)\right]$ is the traveling time cost and traveling distance cost of order-serving flow. The second item $\sum_{(I,J)\in\mathcal{E}_R} \left[x^{Rep}_{(I,J)} \left(V^T T_{(i,j)} + V^D D_{(i,j)}\right)\right]$ is the traveling time cost and traveling distance cost of repositioning flow. The traveling cost includes both the traveling time and distance cost for order-serving and repositioning flow. The third term $\sum_{(I,J)\in\mathcal{E}_C} \left[x^{Rep}_{(I,J)} \left(c_J - c_I\right) \pi^{CS}_{n,t}\right]$ is the charging cost of charging flow. $\pi^{CS}_{n,t}$ is the charging price for each EVCS as issued by CSOs.

Constraint (2) and (3) indicate order-serving and repositioning flows satisfying continuity constraints. In other words, order-serving and repositioning flows entering a node must exit the same node at the same time. An example is provided in the Appendix A to illustrate the continuity condition and flow state transition among order-serving, repositioning, and charging flow as shown in Fig. 16. Constraints (4) represent the start-service flow and end-service flow fulfilling the traveling requests. Constraint (5) is the SoC constraint which indicates that when the SoC of vehicles in node i does not meet the required SoC of road

c_{ij}, both the serving flow and the repositioning flow are zero. Constraints (6) and (7) represent charge limits in EVCS n and road limits in road ij, respectively. Based on the above settings, charging loads $p_{n,t}^{EV}$ can be determined by AMoD operator through the following function.

$$p_{n,t}^{EV} = \sum_{(I,J)\in\mathcal{E}_C} \left[x_{(I,J)}^{Rep}\left(c_J - c_I\right)\right] P^C \quad (\forall n, t_I)\,. \tag{8}$$

B. CSO Operation Model

Without loss of generality, we assume there are multiple profit-driven CSOs providing charging services for the AMoD system. Each CSO owns multiple EVCSs at different locations. The operation objective of a CSO is to maximize the accumulated profit over the entire operation horizon. The power procurement cost is assumed to be determined by the distributed locational marginal price (DLMP) from the distribution system and charging demand. This paper assumes that all the charging demand comes from the AMoD system as the target is to investigate the interactive pattern between charging network and AMoD system.

Each strategic CSO determines its charging pricing strategy to maximize the profit, as shown in the following.

$$\max_{\Xi_2} \quad \rho_m = \sum_{n_m\in N_m,t} \pi_{n_m,t}^{CSO} p_{n_m,t}^{EV}/\eta - \alpha_{n,t} p_{n,t}^{grid}, \tag{9}$$

$$\Xi_2 = \left\{\pi_{n_m,t}^{CSO}, p_{n,t}^{grid}\right\}$$

$$\text{s.t.} \qquad p_{n,t}^{grid} = \frac{p_{n_m,t}^{EV}}{\eta}(\forall n, t), \tag{10}$$

$$\sum_{n_m\in N_m,t} \left(\pi_{n_m,t}^{CSO} p_{n_m,t}^{EV}/\eta - \alpha_{n,t} p_{n,t}^{grid}\right) \geq 0\,(\forall m), \tag{11}$$

$$\pi_{n_m}^{CSO.\min} \leq \pi_{n_m,t}^{CSO} \leq \pi_{n_m}^{CSO.\max}(\forall n_m, m, t). \tag{12}$$

The profit functions (9) of CSO m, which includes EVCS n_m consist of two parts, i.e., charging revenue and negative power procurement cost, where $\alpha_{n,t}$ is DLMP from the distribution system, $p_{n_m,t}^{EV}$ is charging load in EVCS $n_m \in N_m$ belonging to CSO m at time t from the AMoD operator model, and η is the chargers' efficiency of each EVCS. The decision variables of CSO operator model are $\left\{\pi_{n_m,t}^{CSO}, p_{n,t}^{grid}\right\}$, which denote the charging price of EVCS $n_m \in N_m$ managed by CSO $m \in M$ at time t and power procurement from the grid at time t. Constraint (10) indicates power balance for each EVCS n at time t. Constraint (11) indicates that the accumulated profit of each CSO is non-negative. Constraints (12) limit the upper bound and the lower bound the charging prices.

It shall be mentioned that $p_{n_m,t}$ is affected by the price responsive behavior of AMoD system and the pricing strategy of competitive CSOs.

C. Non-cooperative Stackelberg Game

A non-cooperative Stackelberg game formally studies the sequential decision-making processes between CSOs and AMoD operator follower and the non-cooperative interdependence among CSOs [30]. Here, we formulate a non-cooperative Stackelberg game, where the CSOs are the leaders and the AMoD operator is the follower, to capture the interaction between the CSOs and the AMoD operator. In this non-cooperative Stackelberg game, the CSOs play a Nash game with each other to set charging prices. Multiple comparative examples in Section IV are applied to verify the two equilibriums of the non-cooperative Stackelberg game (the Nash equilibrium between CSOs and the Stackelberg equilibrium among CSOs and AMoD fleet operator). The non-cooperative Stackelberg game is formally defined by its strategic form as

$$\Gamma = \left\{ \begin{array}{l} \left(M \cup \{AMoD\}\right), \left\{\pi_m^{CSO}\right\}|_{m\in M}\,, \\ p^{EV}, \{\rho_m\}|_{m\in M}\,, C \end{array} \right\} \tag{13}$$

which consists of the following components.

(1) *Players:* CSOs (leaders) in set M and AMoD operator (follower);
(2) *Strategy sets of players:* The union of feasible strategy sets $\left\{\pi_m^{CSO}\right\}|_{m\in M}$ of all CSOs and p^{EV} of AMoD operator;
(3) *Payoff functions of players:* The profit function $\{\rho_m\}|_{m\in M}$ of each CSO is explained in (9). C is the cost function of AMoD operator, as explained in (1), that captures its total cost including order-serving, repositioning and charging cost.

The solution of the proposed game includes a Stackelberg equilibrium at which the CSOs determine their pricing strategies considering the response of AMoD operator and a Nash equilibrium among CSOs. At this equilibrium point, neither CSO nor AMoD operator can benefit by unilaterally changing their strategy.

Defination 1: Consider the game Γ defined in (13), where $\{\rho_m\}|_{m\in M}$ and C are determined by (9) and (1), respectively. A set of strategies $\left\{\pi_m^{CSO*}\right\}|_{m\in M}, p^{EV*}$ constitutes an equilibrium of this game, if and only if it satisfies the following set of inequalities [31] [32]:

$$\begin{aligned} &\rho_m\left(\pi_m^{CSO*}, \pi_{-m}^{CSO*}, p^{EV*}\right) \\ &\geq \rho_m\left(\pi_m^{CSO}, \pi_{-m}^{CSO*}, p^{EV*}\right)(\forall m), \end{aligned} \tag{14}$$

$$C\left(\pi_m^{CSO*}, p^{EV*}\right) \leq C\left(\pi_m^{CSO*}, p^{EV}\right), \tag{15}$$

where π_m^{CSO*} is the charging prices of CSO m, and π_{-m}^{CSO*} denotes the charging prices of other CSOs. Inequality (14) indicates that no CSO can improve its cost by choosing other strategies rather than π_m^{CSO*}. Inequality (15) shows that AMoD operator cannot decrease its cost by unilaterally deviating from the optimal charging strategy p^{EV*} if responding to the pricing strategies by CSOs.

D. EPEC Formulation

In this paper, each CSO is assumed to have full knowledge of AMoD operator's price responsive patterns. Considering the information asymmetries between CSO and AMoDs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3131804, IEEE Transactions on Smart Grid

operator, the interactive behavior between charging network and AMoD system is formulated as a bi-level optimization problem, as shown in the following.

$$\rho_m = \max_{\Xi_2} \sum_{n_m \in N_m, t} \pi^{CSO}_{n_m,t} p^{EV}_{n_m,t}/\eta - \alpha_{n,t} p^{grid}_{n,t} (\forall m), \quad (16)$$

$$\text{s.t.} \qquad \text{constraints (10)-(12)}, \quad (17)$$

$$\min_{\Xi_1} \begin{bmatrix} \sum_{(I,J)\in\mathcal{E}_R,k} \left[x^{Ser}_{k,(I,J)} \left(V^T T_{(i,j)} + V^D D_{(i,j)} \right) \right] \\ + \sum_{(I,J)\in\mathcal{E}_R} \left[x^{Rep}_{(I,J)} \left(V^T T_{(i,j)} + V^D D_{(i,j)} \right) \right] \\ + \sum_{(I,J)\in\mathcal{E}_C} \left[x^{Cha}_{(I,J)} \left(c_J - c_I \right) P^C \pi^{CS}_{n,t} \right] \end{bmatrix}, \quad (18)$$

$$\text{s.t.} \qquad \text{constraints (2)-(7)}. \quad (19)$$

In the above formulation, (16)-(17) represent the upper-level model for each CSO, and (18)-(19) are the lower-level model for AMoD operator. Meanwhile, the competitive pricing behaviors of CSOs are modeled as a Nash game. Combining those two types of interactive features, we can model the proposed non-cooperative Stackelberg game as following EPEC model:

$$\left\{ \begin{array}{c} \max_{\mathrm{X}_m} \rho_m \left(\mathrm{X}_m \right) \\ \mathcal{H}_m \left(\mathrm{X}_m, \mathrm{X}^*_{-m} \right) = 0 \\ G_m \left(\mathrm{X}_m, \mathrm{X}^*_{-m} \right) \geq 0 \end{array} \right\} \quad (\forall m \in M). \quad (20)$$

Here, function $\rho_m \left(\mathrm{X}_m \right)$ is the profit of CSO m, which is similar to function (9). The decision variables X_m of CSO m are $\{\Xi_1 \cup \Xi_2 \cup \kappa\}$, which is a union set of the decision variable in AMoD system operation model, CSO operation model, and the dual variables κ for AMoD system operation model as $\left\{ \kappa^{SF}_{k,I}, \kappa^{RF}_{I}, \kappa^{SO}_{k}, \kappa^{SD}_{k}, \kappa^{SS}_{(i_I,i_J),t_I}, \kappa^{RS}_{(i_I,i_J),t_I} \kappa^{CR}_{n}, \kappa^{RR}_{(i_I,i_J),t_I} \right\}$. Furthermore, vectors $\mathcal{H}_m \left(\mathrm{X}_m, \mathrm{X}^*_{-m} \right)$ and $\mathcal{G}_m \left(\mathrm{X}_m, \mathrm{X}^*_{-m} \right)$ are equality and inequality constraints of models for CSO m, respectively. Vector X^*_{-m} is the optimal strategies of all other CSOs. Considering that the AMoD operation model has a unique maximum with given prices from CSOs, the non-cooperative Stackelberg game possesses an equilibrium solution if the game among the CSOs admits a Nash equilibrium [30]. To verify Nash equilibrium condition that no individual CSO has a profitable unilateral deviation based on digonalization method [33], $\{\rho^*_m\}|_{m\in M}$ is defined as the objective function value of each CSO operation model solved by the EPEC model. $\widetilde{\rho}_m$ is also defined as the objective function value obtained by solving the bi-level model of CSO m while fixing the pricing decisions of other CSOs. For all CSOs, if the following condition is satisfied

$$\widetilde{\rho}_m \leq \rho^*_m (\forall m), \quad (21)$$

no incremental profit is obtained for any CSO. With this verification process, we can check if the achieved solution is a Nash equilibrium point.

It shall be noted that the EPEC formulation of the proposed non-cooperative Stackelberg equilibrium framework introduces modeling and computational challenges. One of the challenges is due to the non-convexities and non-linearity induced by the Karush-Kuhn-Tucker conditions of the AMoD system operation model. Therefore, it presents computational challenges to solve the proposed equilibrium problem analytically. The other challenge is that solving an equilibrium by analytical approaches requires the information of states and decisions from all CSOs. Such an approach requires each CSO have full knowledge of competitors' strategies, which could be a problematic assumption in real-world applications.

III. Solution Approach: Deep Reinforcement Learning

To address the challenges of solving the proposed non-cooperative Stackelberg equilibrium model, a multi-agent DRL framework based on the SAC algorithm is developed, as illustrated in Fig. 2.

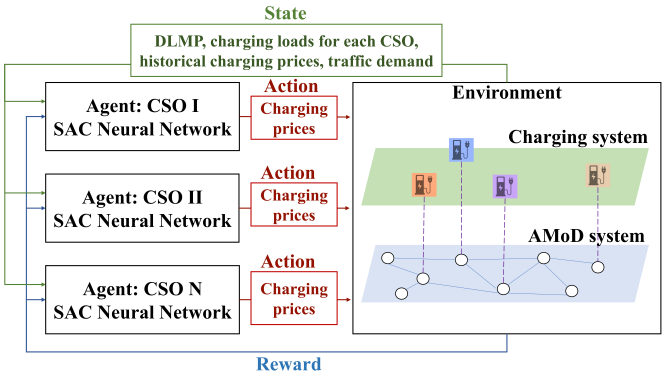


Fig. 2. DRL-based solution framework.

A. DRL-based Solution Framework

We consider a pricing scheme in a day-ahead manner so the AMoD operator can get those price signals in advance. It should be mentioned that the pricing scheme can be adjusted to real-time pricing, but in that case, the responsive model of AMoD operator also needs to be reformulated as receding horizon one [29]. In the proposed framework, each CSO makes pricing decisions in consecutive days $d = 1, 2, \cdots, D$ (called an episode). For each day d, each CSO sets 24-hour prices for the EVCSs. It requires the predetermined DLMP, and the predicted traffic demand, and the history of other CSOs' pricing decisions. The objective of a CSO's pricing decision is to maximize its total profit in an episode. When executing a pricing decision, each CSO can only observe its state without the other CSOs' states and decisions at that moment [34]. Therefore, we consider each CSO as an agent and model the pricing problem as a partially observable Markov decision process in a fully competitive setting. The significant components are defined as:

(1) *Agent:* Each strategic CSO m constitutes the agent.
(2) *Environment:* the environment is represented by the vehicle dispatch and charging management problem

carried out by the AMoD operator, formulated in the optimization problem (1)-(8). Note that the settings of the multi-agent DRL environment, such as AMoD system in the current problem, are set as the same as that in EPEC model.

(3) *State:* For each CSO m, its state variable $s_{m,d}$ can be specified using a set of exogenous attributes and a set of endogeneous attributes [35]. The current timestamp d and DLMP from DSO $[\lambda_{n_m,t}]_{t\in\mathcal{T}_d}$ are included in the exogenous attributes, which is the set of external features of the problem. The charging loads from AMoD system $[p_{n_m,t}]_{t\in\mathcal{T}_d}$ and historical charging prices of all CSOs in the previous day $[\pi_{m,t}]_{m\in M,t\in\mathcal{T}_{d-1}}$, where $\mathcal{T}_d$ denotes all time intervals in day d, are included in the endogeneous attributes, which serves as a feedback signal regarding the influence of its strategic prices on the state of the environment.

(4) *Action:* The action variable of CSO m is the charging prices of EVCSs it manages, i.e., $a_{m,d} = \left[\pi^{CSO}_{n_m,t}\right]_{t\in\mathcal{T}_d}$.

(5) *Reward and return* : The reward of CSO m is

$$r_{m,d} = \sum_{n_m\in N_m, t\in\mathcal{T}_d} \pi^{CSO}_{n_m,t_d} p_{n_m,t_d}/\eta - \alpha_{n,t_d} p^{grid}_{n,t_d}. \quad (22)$$

Then, return $R_{m,d}$ is defined as the cumulative discounted reward of CSO m from time step t until the end of the episode, $R_{m,d)} = r_{m,d} + \gamma r_{m,d+1} + \cdots + \gamma^{D-1-d} r_{m,D-1}$ where discount factor $\gamma \in [0,1]$ reflects the time value of money (the closer γ is to 1, the more important are future rewards), D denotes all the day in one training episode. The state variables and rewards interacting with the multi-agent DRL agents usually need to be normalized (e.g., normalized to the $0 \sim 1$ range), which is helpful for training the neural network [36].

B. Soft Actor-Critic (SAC) Algorithm

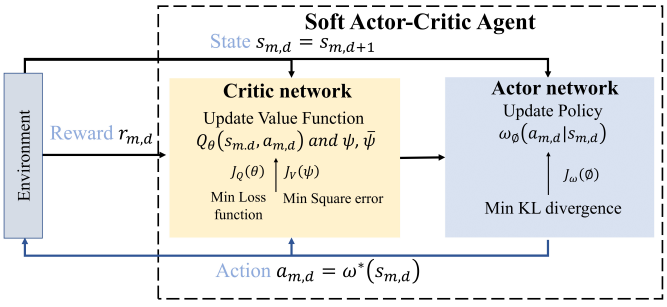


Fig. 3. Learning process of the SAC agent

The traditional on-policy DRL algorithms such as Proximal Policy Optimization (PPO) are sample inefficient since new samples must be generated at each gradient step. Although off-policy policy gradient algorithms, such as Deep Deterministic Policy Gradient (DDPG) [37], were developed to improve sample efficiency, they are often brittle concerning their hyperparameters resulting in poor convergence performance. To address these challenges, the off-policy maximum-entropy deep RL algorithm, SAC [28], is adopted to provide robust and sample-efficient learning, which achieves a better performance. The training process of the SAC agent is as follows. For simplicity, we omit the subscript m for all variables in this section. As shown in Fig. 3, the SAC architecture consists of the actor part and critic part:

$$J_V(\psi) = \mathbb{E}_{s_d}[\frac{1}{2}(V_\psi(s_d) - \mathbb{E}_{a_t}[Q_\theta(s_d, a_d) - \log\omega_\phi(a_d \mid s_d)])^2]. \quad (23)$$

Critic: The critic part contains two value functions $V(S_d)$ parameterized by $\psi, \bar{\psi}$, and two Q-value functions $Q(s_d, a_d)$ parameterized by θ_1, θ_2. Note that similar to the DDPG [38], a target value function $V_{\bar{\psi}}(s_d)$ is not trainable but softly updated to $V_\psi(s_d)$ gradually, $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$, while $V_\psi(s_d)$ is trained to minimize the square error:

$$\nabla_\psi J_V(\psi) = \nabla_\psi V_\psi(s_d)(V_\psi(s_d) - Q_\theta(s_d, a_d) + \log\omega_\phi(a_d \mid s_d))^2. \quad (24)$$

where $-\log\omega_\phi(a_d \mid s_d)$ indicates the policy entropy. The gradient of $J_V(\psi)$ can be calculated as

Similar to the double DQN [39], two Q-value functions are adopted here. When updating action a_d, the minimum one would be picked up to prevent overestimation. The same loss function $J_Q(\theta)$ for θ_1 and θ_2 is shown as follows:

$$J_Q(\theta) = \mathbb{E}_{s_d,a_d}[\frac{1}{2}(Q_\theta(s_d, a_d) - (r(s_d, a_d) + \gamma\mathbb{E}_{s_{d+1}}[V_{\bar{\psi}}(s_{d+1})]))^2]. \quad (25)$$

Then we get the gradient of $J_Q(\theta)$:

$$\nabla_\theta J_Q = \nabla_\theta Q_\theta(s_d, a_d)(Q_\theta(s_d, a_d) - r(s_d, a_d) - \gamma V_{\bar{\psi}}(s_{d+1}t)). \quad (26)$$

Actor: The actor part is the policy network $\omega(a_d \mid s_d)$ parameterized with ϕ. To update the parameters ϕ, the minimization of KL divergence between the policy and the exponential of the Q-function is adopted. The objective function is

$$J_\omega(\phi) = \mathbb{E}_{s_t}\left[D_{\mathrm{KL}}\left(\omega_\phi(\cdot \mid s_d) \,\|\, \frac{\exp(Q_\theta(s_d, \cdot))}{z_\theta(s_d)}\right)\right]. \quad (27)$$

To make it differentiable, a reparameterization trick is adopted, i.e., using the randomness of random noise ϵ_d following Gaussian distribution to substitute the randomness sampling. Then, the action a_d including mean and variance is $a_d = f_\phi(\epsilon_d; s_d)$. In this way, the reconstructed objective function is differentiable:

$$J_\omega(\phi) = \mathbb{E}_{s_d}[\log\omega_\phi(f_\phi(\epsilon_d; s_d) \mid s_d) - Q_\theta(s_d, f_\phi(\epsilon_d; s_d))]. \quad (28)$$

We can get the gradient of $J_\omega(\phi)$ as

$$\hat{\nabla}_\phi J_\omega(\phi) = \nabla_\phi \log\omega_\phi(a_d \mid s_d) + (\nabla_{a_t}\log\omega_\phi(a_d \mid s_d) - \nabla_{a_d} Q(s_d, a_d))\nabla_\phi f_\phi(\epsilon_d; s_d). \quad (29)$$

C. Workflow for Multi-CSO SAC Algorithm

We assume that multiple CSOs make charging pricing strategies simultaneously. For each day d, each CSO m observes the state $s_{m,d}$ from the DSO and AMoD systems and select the action $a_{m,d}$ with mean and variance according to its policy ω_{ϕ_m}. Next, the CSO receives the reward $r_{m,d}$ and observes a new state $s_{m,d+1}$. Based on these experiences stored in the replay buffer, we can calculate the gradients of all objective functions (i.e., Eqs. (24), (26), and (29)) to update the neural network parameters. Then, the interaction process is looped until the rewards of the agents converge. After converge, to guarantee non-negative profit of each CSO, we verify the result using constraint (11). Finally, the proposed multi-CSO SAC algorithm will achieve the Nash equilibrium point. Algorithm 1 shows the training process of the multi-CSO SAC algorithm.

Algorithm 1 Multi-CSO SAC Algorithm

for each CSO m **do**
  Initialize replay buffer D_m
  Initial parameter vectors $\psi_m, \bar{\psi}_m, \theta_{m,1}, \theta_{m,2}, \phi_m$
end for
repeat
  for each environment step **do**
    $a_{m,d} \sim \omega_{\phi_m}\left(a_{m,d} \mid s_{m,d}\right)$ for each CSO m;
    $p_{n,t} \leftarrow$ Eqs. $(1)-(8)$
    for each CSO m **do** in parallel
      $r_{m,d} \leftarrow$ Eq. (22)
      Observe $s_{m,d+1}$
      $\mathcal{D}_m \leftarrow \mathcal{D}_m \cup \{s_{m,d}, a_{m,d}, r_{m,d}, S_{m,d+1}\}$
    end for
  end for
  for each gradient step **do**
    for each CSO m **do** in parallel
      $\psi_m \leftarrow \psi_m - \eta_V \hat{\nabla}_{\psi_m} J_V\left(\psi_m\right)$
      $\theta_{m,i} \leftarrow \theta_{m,i} - \eta_Q \hat{\nabla}_{\theta_{m,i}} J_Q\left(\theta_{m,i}\right)$
        for $i \in \{1,2\}$
      $\phi_m \leftarrow \phi_m - \eta_\pi \hat{\nabla}_{\phi_m} J_\pi\left(\phi_m\right)$
      $\bar{\psi}_m \leftarrow \tau \psi_m + (1-\tau)\bar{\psi}_m$
    end for
  end for
until convergence
non-negative reward verification using (11)

IV. Numerical Experiments

In this section, the simulation results for the non-cooperative Stackelberg equilibrium model based on the Dallas-Fort Worth metroplex data are presented and analyzed. All the experimental simulations are run on a computer with 4 cores Intel Core i7 and 8 GB memory. The DRL-based framework is constructed by the PyTorch in Python.

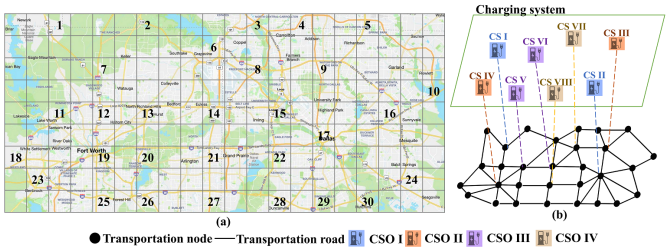


Fig. 4. Charging and transportation systems in Dallas-Fort Worth metroplex. (a) 180-grid transportation system of Dallas-Fort Worth. (b) Coupled charging system and transportation network.

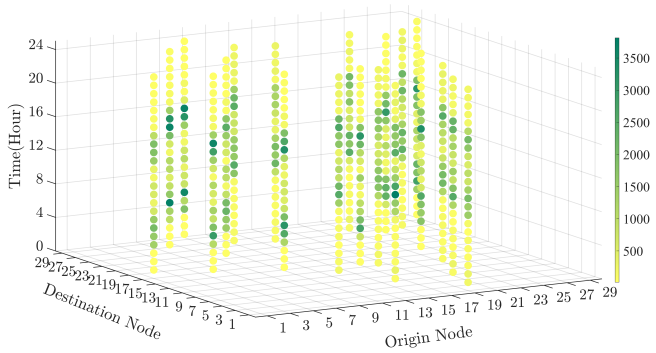


Fig. 5. Daily traveling requests profile.

A. Parameter Settings

The topology of the transportation system in the Dallas-Fort Worth metroplex is shown in Fig. 4, which is modified based on [29]. The targeting area is divided into 180 grids with 5 times 5 kilometer resolution, and the corresponding road network contains 30 nodes and 116 roads as simplified from OpenStreetMap [40]. The charging infrastructures in each node are aggregated, resulting in 8 EVCSs located at Node 5, 7, 8, 11, 17, 19, 24, and 27 in the transportation network. The charging efficiency η of each charging station is set as 92%, based on [41]. Four competitive CSOs are operating those 8 EVCSs, as shown in Fig. 4(b). The EVCS parameters are listed in Table I. The daily traveling requests for AMoD are assumed as 390,000 based on [29]. The commuters' value of time and distance is set equal to \$4.40/h and \$0.1/km, respectively [29]. The BAIC EV 200 is used as an exemplary EV model in the revised manuscript. The EV battery capacity is 3.3 kW/30.4 kWh that can support roughly 200 km driving distance. Fig. 5 shows the daily traveling requests profiles. The AMoD fleet consists of 150,000 vehicles, i.e., 1 AMoD fleet for every 2.6 customers, similar to [11]. To represent the possibility that vehicles might not begin the day fully charged, each EV starts the time episode with a 50% battery charge and must have the same level of charge at the end of each day. The time interval for pricing is 1 hour which is consistent with the wholesale power market settlement interval. In the SAC algorithm, the pricing time interval is set to 24 hours (one day), while each episode is considered 7 days to address the day-to-day

coupling. The other parameters in the training process are set as follows: the number of episodes is 4000, the capacity of replay buffer $D = 5 \times 10^5$, and the discount factor $\gamma = 0.9$.

TABLE I
EVCS PARAMETERS.

Name	Symbol	Value
The vehicle capacity [29]	C_n^c	6000 vehicles
The chargers' efficiency [41]	η	0.92
Upper bound of charging prices	$\pi_n^{CS.\max}$	150 \$/MWh
Lower bound of charging prices	$\pi_n^{CS.\min}$	0 \$/MWh

B. Result Analysis

1) Base Case Result

As stated in section III, the charging pricing strategies of CSO agents result from the learning process in an interactive environment by trials and errors. Here, a set of convergent charging prices during one day are selected as the result analysis, as shown in Fig. 6. From the results, one can see that charging prices and charging loads are negatively correlated. For instance, the charging loads in CSO II are high from 13:00 to 19:00, while the following prices are lower than other time intervals. Besides, it is also significant for charging management of charging prices on a spatial scale. For instance, compared with Fig. 6(b), more vehicles are navigated to EVCS IV than EVCS II due to the lower prices in EVCS IV from 6:00 to 7:00. Fig. 7 shows the distribution of transportation flows in the Dallas-Fort Worth metroplex from 6:00 to 7:00. It can be observed that the flow distribution is affected by the charging pricing strategy. For example, AMoD operator prefer to choose "Node 22 - Node 9" rather than "Node 22 - Node 17" due to the lower prices in Node 21.

In the transportation system, AMoD operator reposition the vehicles to fulfill future traveling requests so that there is an opposite trend between order-serving flow and repositioning flow, especially when there are not enough vehicles for requests. Fig. 8 shows the number of repositioning fleets for Destination Nodes 4, 7, and 17. Destination Nodes 4 and 7 for repositioning fleets have more flows during the early time, following Origin Nodes 4, 7 for requests. To reach a 50% level of charge for fleets and the minimum number for each bus at the end of the day, the repositioning flow for each node looks high from 20:00 to 23:00. To further present the network flow, Fig. 9 depicts the number of the order-serving flow and repositioning flow of Node 17 (EVCS II) from 6:00 to 7:00, which are mainly constrained by constraints (2) and (3). In Fig. 9(b) and (c), one can track the flow in Node 17. To ensure that the SOC can cover the travel, the vehicles that flow into the node at the end of the service generally reduce SOC, and the vehicles that begin to flow out of the node generally increase SOC. For instance, there are 207 vehicles in Node 17 when SOC equals 1, while 497 vehicles out of Node 17 when SOC equals 9.

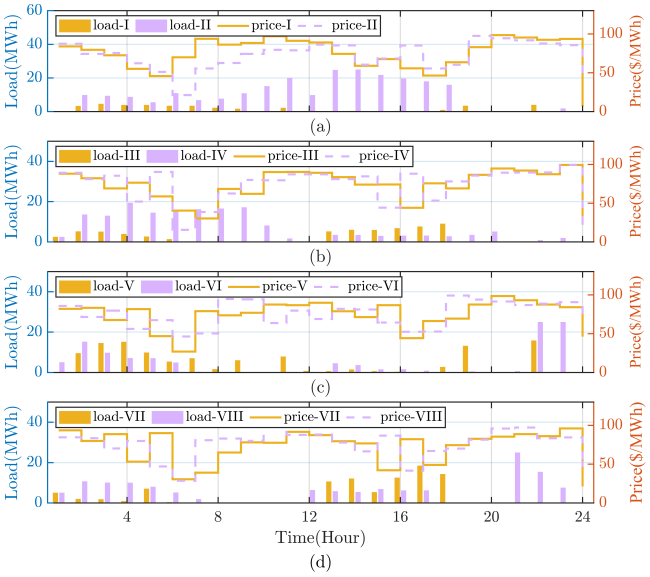


Fig. 6. Charging load and charge prices of (a) CSO I (EVCS I-II), (b) CSO II (EVCS III-IV), (c) CSO III (EVCS V-VI) and (d) CSO IV (EVCS VII-VIII). Load-N price-N in legend represent charging loads and charging prices of EVCS N.

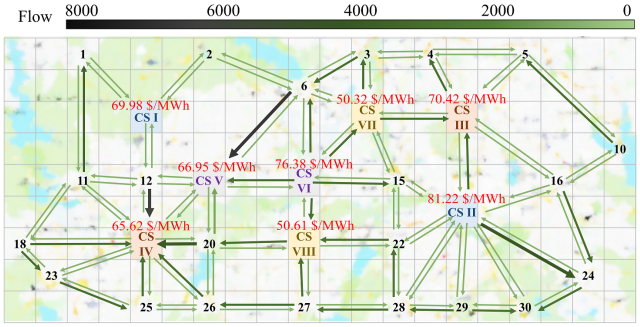


Fig. 7. Distribution of transportation flows in Dallas-Fort Worth metroplex from 6:00 to 7:00.

To illustrate the traceability of the order-serving flows for each travel demand k, Fig. 10 represents the three sets of order-serving flows for node 17 departing at 6 o'clock. Fig. 10(a) and (b) represent three travel demands from node 17 with a departure time of 6:00 in time and SOC, respectively. 10 (c) is a projection of Fig. 10(a) and (b), clearly showing the trajectory of the traffic flow in a geographical map.

2) Comparison with EPEC

We compare and analyze the results from the EPEC method and proposed SAC algorithm. The experiments for the SAC approach are repeated with ten random seeds to prevent contingency. The final convergence result of the reward of each CSO is shown in Fig. 11. The error ranges of the results of the four CSOs relative to the EPEC results are (-1.68%, 1.49%), (-1.49%, 1.97%), (-1.59%, 2.06%) and (-1.48%, 2.27%), respectively. According to reference [25] and [42], the results prove the accuracy and stability of the SAC algorithm with the acceptable error range. In order to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3131804, IEEE Transactions on Smart Grid

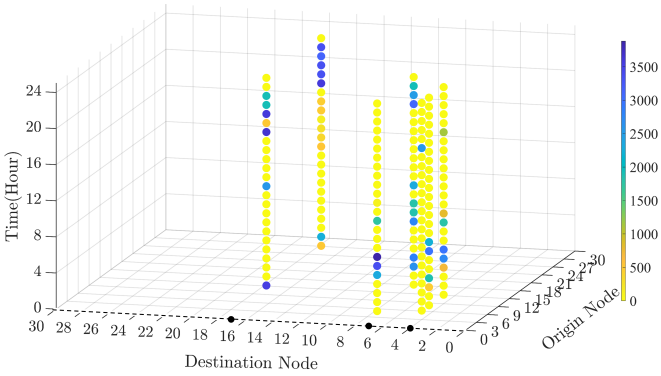


Fig. 8. The number of repositioning AMoD fleets for Destination Node 4, 7 and 17.

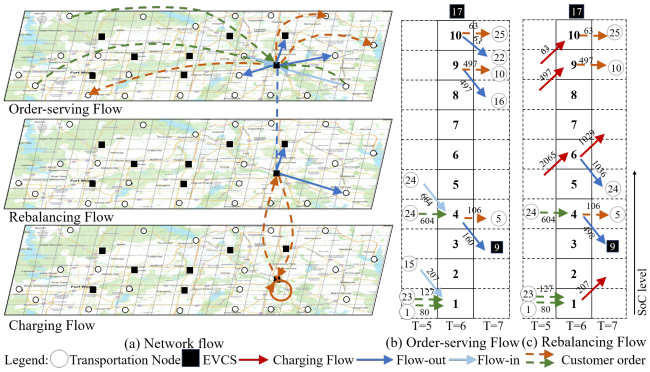


Fig. 9. The network flow of Node 17 (EVCS II) from 6:00 to 7:00.

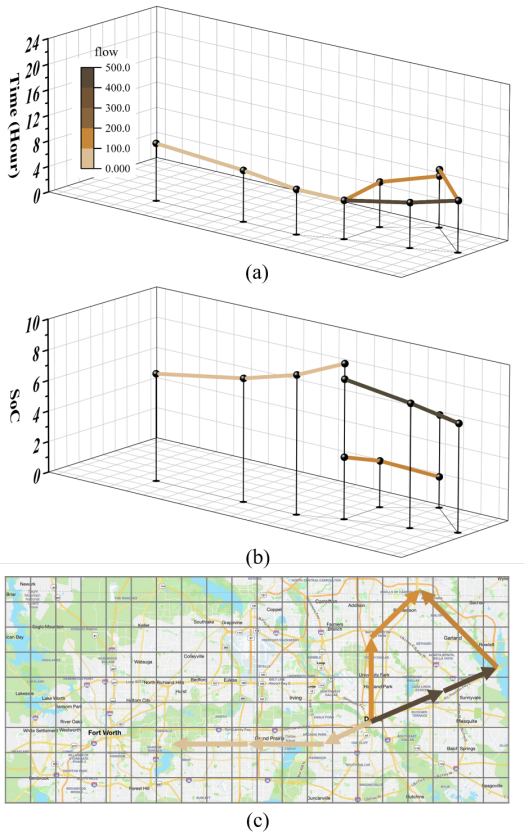


Fig. 10. Illustration of traceability. (a) Time trajectory. (b) SoC trajectory. (c) Geographical trajectory.

verify the Nash equilibrium, we adopted the diagonalization method in [30]. Table II verifies that the solution obtained in Fig. 6 is a Nash equilibrium point. It can be seen that The objective function values ρ_1^* of each CSO solved by the EPEC model are both larger than the objective function value $\widetilde{\rho_m}$ obtained by solving the MPEC model with fixed charging pricing strategies of other CSOs obtained by the EPEC model. It shows that the solution conforms to the Nash equilibrium condition that no individual CSO has a profitable unilateral deviation, and the solution of the EPEC model is Nash equilibrium.

TABLE II
VERIFICATION OF NASH EQUILIBRIUM.

CSO I (k$)		CSO II (k$)		CSO III (k$)		CSO IV (k$)	
ρ_1^*	$\widetilde{\rho}_1$	ρ_2^*	$\widetilde{\rho}_2$	ρ_3^*	$\widetilde{\rho}_3$	ρ_4^*	$\widetilde{\rho}_4$
11.76	10.21	8.67	8.21	10.61	9.54	11.39	10.99

3) Comparison with Centralized Approach

Compared to centralized approach, there is a loss of efficiency in decentralized approach. We add a benchmark for centralized approach and solve it with two methods, i.e., centralized approach based mathematical programming with equilibrium constraints (C-MPEC) and SAC (C-SAC). To illustrate the difference in performance, we establish an efficiency loss index [43] to quantify the impact of adopting

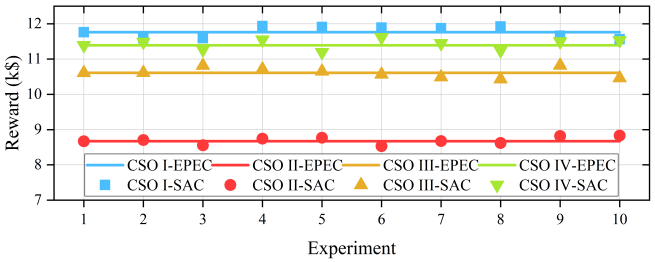


Fig. 11. The final convergence value of rewards (the lines represent the solutions of EPEC and the dots represent the solution of ten experiments with different random seeds for SAC algorithm).

decentralized approaches, i.e., decentralized approach based EPEC (D-EPEC) and SAC (D-SAC). The results in Table III demonstrate that there is a slight performance difference between the centralized and decentralized approaches, e.g. 0.28% for D-EPEC in contrast to C-MPEC and 0.24% for D-SAC in contrast to C-SAC.

TABLE III
COMPARATIVE RESULT OF EFFICIENCY LOSSES.

Name	C-MPEC	C-SAC	D-EPEC	D-SAC
Total/Joint profit (k$)	42.51	42.50	42.39	42.40
Efficiency loss index	-	-	0.28%	0.24%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3131804, IEEE Transactions on Smart Grid

C. Sensitivity Analyses

1) The Impact on the Operation Efficiency of the Coupled Charging System and AMoD System

To demonstrate the impact of the proposed pricing scheme on the operation efficiency of the coupled charging system and AMoD system, we compare the proposed pricing strategy as a base case with three intuitive benchmark cases, and the results are reported as follows.

Case A: no spatial difference for all charging prices.

Case B: no temporal difference for all charging prices.

Case C: no spatial-temporal difference for all charging prices.

TABLE IV
ECONOMIC PERFORMANCE OF AMOD SYSTEM AND CSO SYSTEM.

	Cost of AMoD system			Profit of CSO system	
	Charging (k$)	Traveling (k$)	Total (k$)	Cost (k$)	Profit (k$)
Base case	334.77	32.15	366.92	292.34	42.43
Case A	357.95	28.28	386.23	329.51	28.44
Case B	367.19	30.52	397.71	339.77	27.42
Case C	377.17	27.16	404.33	372.05	5.12

Table IV shows the economic performance of AMoD system's weekly charging revenue in different benchmarks. It can be observed that adopting the proposed pricing strategy in the base case leads to lower charging costs and higher traveling costs for the AMoD system. This is because the prices in other benchmarks cannot include the spatial-temporal coordination for AMoD operator's order-serving, repositioning, and charging. In other words, AMoD system might make less-distance order-serving or repositioning decisions without considering the spatial-temporal distribution of charging prices. One can also see that adopting other benchmark pricing strategies leads to more energy procurement costs and fewer profits for the CSOs. The results demonstrate that the proposed spatial-temporal charging pricing strategy improves the operation efficiency of the coupled charging system and AMoD system.

2) The Impact on the Competitive Performance among CSOs

To illustrate the competitive performance of the proposed charging pricing-based equilibrium strategy, two more benchmarks are investigated.

Case D: CSO I, CSO II, and CSO III adopt the proposed scheme, while CSO IV utilizes a fixed time-varying price policy as the base case.

Case E: CSO I, CSO II, and CSO III adopt the proposed time-varying scheme while CSO IV utilizes the proposed pricing scheme but with the additional time-invariant constraint.

TABLE V
WEEKLY PROFITS OF CSO SYSTEM.

	CSO I (k$)	CSO II (k$)	CSO III (k$)	CSO IV (k$)	Total (k$)
Base case	11.76	8.67	10.61	11.39	42.43
Case D	12.52	10.32	11.31	6.59	40.74
Case E	12.05	9.05	10.98	8.73	40.81

It can be observed in Table V that the revenues of CSO I, CSO II, and CSO III in Case D and Case E are higher than the result of the base case, while the revenue of CSO IV in Case D and Case E is lower than the result of the base case. Fig. 12 shows the charging demand management results in Case D. It can be observed that CSO I, CSO II, and CSO III adopting the strategic pricing method attract more charging demand from CSO IV (compared to Fig. 6), resulting in higher revenues for those strategic players.

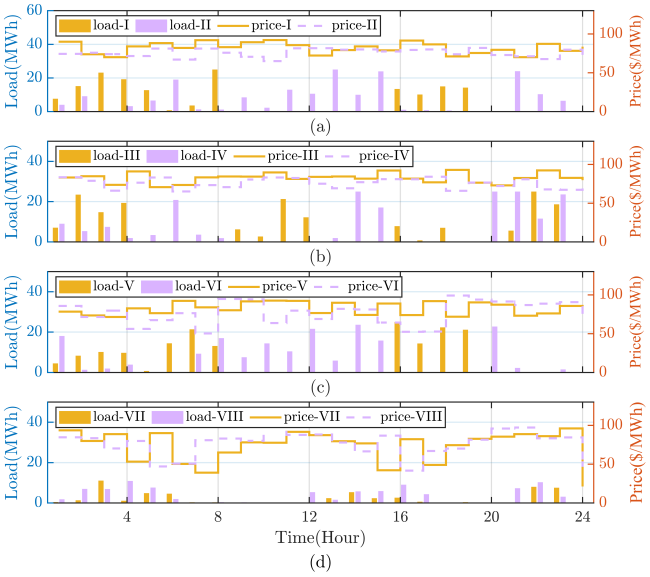


Fig. 12. Charging load and charge prices in Case D. (a) CSO I (EVCS I-II), (b) CSO II (EVCS III-IV), (c) CSO III (EVCS V-VI) and (d) CSO IV (EVCS VII-VIII). Load-N price-N in legend represent charging loads and charging prices of EVCS N.

3) The Impact of Forecasts Quality on the Rewards of CSOs

To explore the effect of the quality of forecasts of travel requests and DLMP parameters on the rewards of each CSOs, four sensitivity analysis cases are provided. In each case, the Normalized Standard Deviation of Day-ahead Forecasting Error of DLMP and travel requests are 0.1, 0.2, 0.3 and 0.4, respectively. The agents sample the DLMP and travel requests for each case based on the Monte Carlo sampling method [44]. Ten repeated experiments are executed, and the error range and variance results with the EPEC results are shown in Table VI. It can be observed that the variance of results increases as the Normalized Standard Deviation of Day-ahead Forecasting Error increases. However, the variance is acceptable even when the Normalized Standard Deviation of Day-ahead Forecasting Error of DLMP and travel requests is 0.4, according to the literature [38][45]. It shows that the neural network is adaptive to environmental states with different error ranges.

D. Computation Performance

To validate the computation performance of the SAC based learning strategy, Fig. 13 depicts the convergence process of the proposed multi-agent SAC algorithm. It shows that the convergence can be obtained after observing and training for about 4000 episodes, which embodies the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSG.2021.3131804, IEEE Transactions on Smart Grid

TABLE VI
SENSITIVITY RESULT IN TEN REPEATED EXPERIMENTAL RESULTS BASED ON DIFFERENT FORECAST QUALITIES.

		Normalized Standard Deviation of Day-ahead Forecasting Error of DLMP and travel requests			
		0.1	0.2	0.3	0.4
CSO I	Error range(%)	(-1.68, 1.49)	(-2.95, 2.51)	(-3.63, 3.40)	(-19.40, 5.00)
	Variance(k^2)	0.021	0.074	0.818	9.793
CSO II	Error range(%)	(-1.49, 1.97)	(-2.07, 3.35)	(-2.10, 4.35)	(-16.63, 20.03)
	Variance(k^2)	0.011	0.018	0.165	2.34
CSO III	Error range(%)	(-1.59, 2.06)	(-1.96, 2.99)	(-4.48, 4.83)	(-19.65, 18.68)
	Variance(k^2)	0.017	0.032	0.174	0.412
CSO IV	Error range(%)	(-1.48, 2.27)	(-2.62, 3.52)	(-4.25, 4.73)	(-20.65, 12.03)
	Variance(k^2)	0.022	0.109	2.14	10.385

desirable convergence properties. Besides, Fig. 14 shows the total weekly return of agents during the training for SAC, PPO, and DDPG. The results show that the proposed model outperforms other state-of-art model-free deep RL algorithms, including the on-policy PPO algorithm and off-policy DDPG algorithm [28]. One can see that the PPO algorithm needs more episodes (about 4000 episodes) than the SAC algorithm due to the lower sampling efficiency. Besides, the DDPG algorithm fails to make any progress without convergence due to its extreme brittleness and hyper-parameter sensitivity [28].

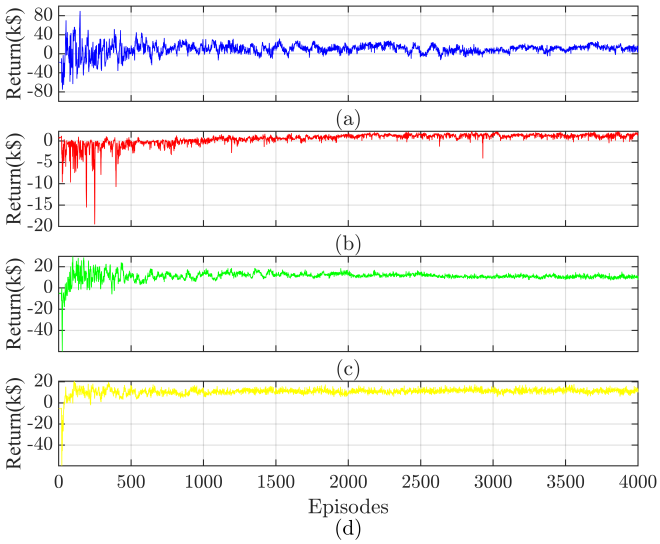


Fig. 13. Episodic return of CSOs (a) CSO I; (b) CSO II; (c) CSO III; (d) CSO IV.

TABLE VII
COMPUTATIONAL PERFORMANCE COMPARISON BETWEEN EPEC AND DRL METHODS.

	EPEC	DRL
Time	156 minutes	$<$ 1 second

It can be observed that the proposed SAC algorithm can obtain converged results within 4000 episodes. With the help of well-trained neutral networks, CSO can determine its optimal charging prices based on non-proprietary information such as DLMP and trip request distribution. In this way, a minimal computational burden is required by the proposed DRL based pricing scheme when determining the charging price. This is illustrated in Table VII by compared the computational performance of the proposed DRL method with the conventional EPEC solution method based on [46]. It can be observed that the proposed DRL method is much faster than the EPEC method. This is expected as the analytical solution method for EPEC needs to deal with the inherent non-convexities and non-linearity. In contrast, the machine learning based solution method can immediately find the optimal mapping between the input parameter and optimal pricing strategy after finishing the offline training of the algorithm. It worth mentioning that this could be an even more important feature if we extend the charging pricing to an online manner in the future.

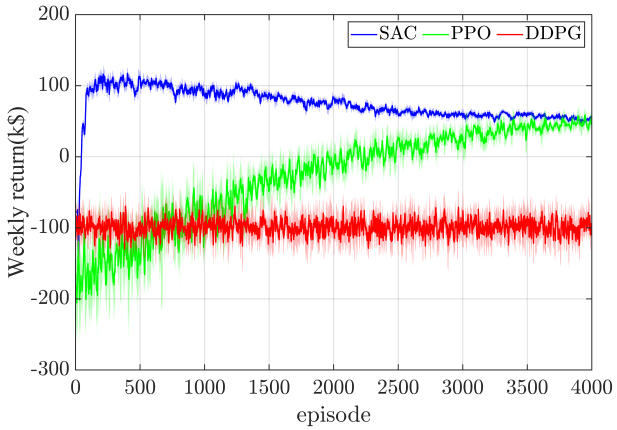


Fig. 14. Training results on model-free deep RL algorithms. The solid curves are the performances of 5 random experiments, and the shaded regions represent the error bounds.

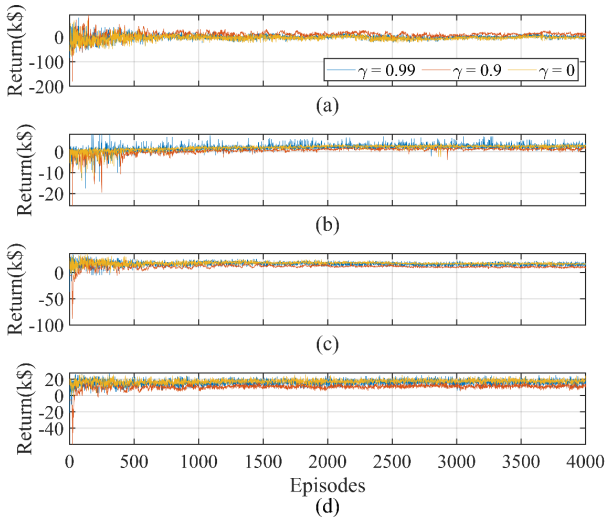


Fig. 15. Episodic return of CSOs with different discount factor of CSO I. (a) CSO I; (b) CSO II; (c) CSO III; (d) CSO IV. (γ is the discount factor)

The key settings of MADRL itself are mainly hyperparam-

eters, such as learning rate and discount factor. However, the setting of hyperparameters usually lacks theoretical support and depends mainly on empirical and experimental results. Some empirical conclusions are 0.01 $\sim$ 0.0001 for the learning rate and 0.8 $\sim$ 0.99 for the discount factor, but still some experiments are needed to tune the hyperparameters for specific problems. We conduct experiments to investigate the effect of the discount factor on the returns of the CSOs. The discount factor for CSO I is set to vary, while the discount factor for CSO II-IV is a constant 0.9. The returns during training can be seen in Fig. 15. A low discount factor can lead agents to over-prioritize immediate returns and become myopic about future returns [47]; however, targeting a high discount factor may lead to instability or divergence in the estimation of the Q-value function, yielding a poor quality policy [45]. Both $\gamma = 0.99$ and $\gamma = 0$ reduce the returns of CSO I, while $\gamma = 0.9$ implies convergence of the more profitable policy. When CSO I uses a less profitable policy, the other CSOs' policies are more competitive thus yielding greater returns. It can be seen that, too large or too small discount factors will significantly reduce the performance of the algorithm. When discount factors equal to 0.9, the proposed method achieves a useful trade-off.

Note that DRL usually requires a lot of data from different scenarios. The scarcity of data for certain scenarios, such as extreme weather, may hinder the performance of DRL model in such scenarios. One solution is to generate data for rare scenarios, e.g., combine other information to predict traffic, grid data in these scenarios and extend existing data using data augmentation techniques. In addition, our model can be easily combined with some techniques to improve training efficiency, such as using Graph convolutional neural networks (GCNs) [48] to improve feature extraction efficiency for traffic network and power grid data, and using transfer learning methods to improve learning adaptability and efficiency for multiple scenarios.

V. Conclusion

In this paper, we investigate a strategic charging pricing scheme for CSOs based on a non-cooperative Stackelberg equilibrium framework while the unique operational characteristics of AMoD system are considered. In the proposed non-cooperative Stackelberg equilibrium framework, the equilibrium studies the pricing competition among multiple CSOs, and explores the nexus between the CSOs and AMoD operator. A MADRL framework based on a SAC algorithm is established to solve the proposed equilibrium framework while privacy-conservation constraints among CSOs are considered. Simulation results of the city-scale real-world Dallas-Fort Worth metroplex verify the effectiveness of the proposed framework. The results demonstrate that the proposed spatial-temporal charging pricing strategy improves the operation efficiency of the integrated charging system and transportation. Also, it verifies the proposed competitive pricing strategies for commercial CSOs outperform other benchmark pricing methods.

Appendix A

A small example is given to illustrate the proposed multi-commodity network flow model. As shown in Fig. 16(a), The topology of the transportation network includes 4 nodes ($N1$, $N2$, $N3$, and $N4$), 10 roads, 1 origin-destination (O-D) pair ($k = (N1, N4, T2, \lambda 8)$ meaning that 8 vehicles are required from $N1$ to $N4$ starting at $T2$), and 1 EVCS located in $N1$. The augmented energy-time-space graph is given in Fig. 16(b), in which the network can be expanded to one with 36 virtual nodes ($I = (N1, C1, T1), \cdots, (N4, C3, T3)$), 60 road links ($\mathcal{E}_R$) and 6 charging process ($\mathcal{E}_C$). It is assumed that the required SoC of one vehicle to pass through each road is 1 p.u., and the required time to pass through each road is 1 p.u. The number of available vehicles in the initial time is 3 vehicles at node $(N1, C2, T1)$, 3 vehicles at node $(N2, C3, T1)$ and 3 vehicles at node $(N3, C3, T1)$. The distance of $(N2, N1)$, $(N3, N1)$, and $(N4, N1)$ are 20km, 23km, 25km.

TABLE VIII
Results of the illustrative example.

	Related symbol	AMoD fleet
Serving flow	$x^{Ser}_{k1,((N1,C3,T2),(N2,C2,T3))}$	3 vehicles
	$x^{Ser}_{k1,((N1,C2,T2),(N2,C1,T3))}$	5 vehicles
Repositioning flow	$x^{Rep}_{k1,((N2,C3,T1),(N1,C2,T2))}$	3 vehicles
	$x^{Rep}_{k1,((N3,C3,T1),(N1,C2,T2))}$	2 vehicles
Charging flow	$x^{Rep}_{k1,((N1,C2,T1),(N1,C3,T2))}$	3 vehicles
Start-service flow	$z^{ori}_{k1,C2,T3}$	3 vehicles
	$z^{ori}_{k1,C2}$	5 vehicles
End-service flow	$z^{des}_{k1,C1,T3}$	3 vehicles
	$z^{des}_{k1,C2}$	5 vehicles

Without vehicle repositioning, only 3 available vehicles exist (charge) in $N1$, which cannot fulfill the travel requests. Therefore, repositioned vehicles from the neighboring nodes ($N2$, $N3$, and/or $N4$) are needed. Table IV shows the results of each network flows, including serving flow, repositioning flow, and charging flow.

(a) Serving flows (green arrows in Fig. 16) visually explain constraints (2). For nodes $(N1, T2, C3)$ and $(N1, T2, C2)$ which are the origin nodes of OD, the constraints can be presented as $z^{ori}_{k1,C3} = x^{Ser}_{k1,((N1,C3,T2),(N2,C2,T3))}$ and $z^{ori}_{k1,C2} = x^{Ser}_{k1,((N1,C2,T2),(N2,C1,T3))}$. For nodes $(N4, T3, C2)$ and $(N4, T3, C1)$ which are the destination nodes of OD, the equations are established as $z^{des}_{k1,C2,T3} = x^{Ser}_{k1,((N1,C3,T2),(N2,C2,T3))}$ and $z^{des}_{k1,C1,T3} = x^{Ser}_{k1,((N1,C2,T2),(N2,C1,T3))}$. Constraints (4) are defined to fulfill the travel request at the origin node and destination node at the same time. For $k = [(N1, N4, T2, \lambda 8)]$, both $N1$ and $N4$ need to satisfy the equation $z^{ori}_{k1,C3} + z^{ori}_{k1,C2} = z^{des}_{k1,C2,T3} + z^{des}_{k1,C1,T3} = 8$.

(b) Repositioning flows (blue arrows in Fig. 16) visually explain constraints (3). For node $(N1, T2, C2)$,

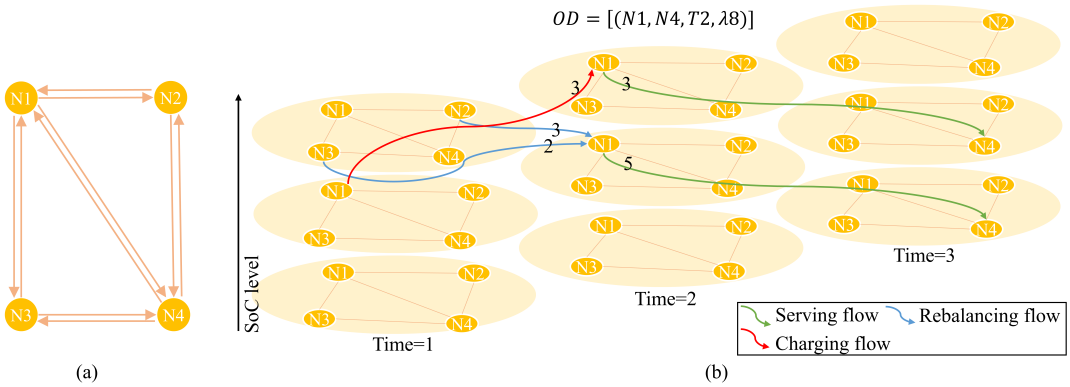


Fig. 16. Diagram of an illustrative example including (a) The topology of the transportation network, and (b) Augmented energy-time-space graph.

which is the starting nodes of k, so the equations can be presented as $x^{Rep}_{((N2,C3,T1),(N1,C2,T2))} + x^{Rep}_{((N3,C3,T1),(N1,C2,T2))} = z^{ori}_{k1,C2}$ which means that the vehicle that ends the dispatch is equal to the vehicle that starts the dispatch. As the low distance cost and shorter time cost, AMoD fleets are repositioned by $(N2, N1)$ first.

(c) Charging flows (Red arrows in Fig. 16) also explain constraints (3). For node $(N1, T2, C3)$, which is the starting nodes of OD, so the equations can be presented as $x^{Rep}_{((N1,C2,T1),(N1,C3,T2))} = z^{ori}_{k1,C3}$ which means that the number of end-charge vehicles is equal to the vehicle that starts the service.

REFERENCES

[1] W. Su, H. Eichi, W. Zeng, and M. Y. Chow, "A survey on the electrification of transportation in a smart grid environment," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 1–10, 2012.

[2] J. Hamari, M. Sjöklint, and A. Ukkonen, "The sharing economy: Why people participate in collaborative consumption," *Journal of the Association for Information Science and Technology*, vol. 67, no. 9, pp. 2047–2059, 2016.

[3] J. Zhao, H. Wang, Y. Liu, Q. Wu, Z. Wang, and Y. Liu, "Coordinated Restoration of Transmission and Distribution System Using Decentralized Scheme," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 3428–3442, 2019.

[4] E. Martin and S. Shaheen, "Impacts of car2go on Vehicle Ownership, Modal Shift, Vehicle Miles Traveled, and Greenhouse Gas Emissions: An Analysis of Five North American Cities," Transportation Sustainability Research Center, UC Berkeley 3, Tech. Rep., 2016. [Online]. Available: http://innovativemobility.org/wp-content/uploads/2016/07/Impactsofcar2go_FiveCities_2016.pdf

[5] F. Ciari, B. Bock, and M. Balmer, "Modeling station-based and free-floating carsharing demand: Test case study for Berlin," *Transportation Research Record*, vol. 2416, pp. 37–47, 2014.

[6] CBINSIGHTS, "CB Insights' technology insights platform." [Online]. Available: https://www.cbinsights.com/research/autonomous-driverless-vehicles-corporations-list/

[7] M. Pavone, S. L. Smith, E. Frazzoli, and D. Rus, "Robotic load balancing for mobility-on-demand systems," *International Journal of Robotics Research*, vol. 31, no. 7, pp. 839–854, 2012.

[8] R. Zhang and M. Pavone, "Control of robotic mobility-on-demand systems: A queueing-theoretical perspective," *International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 186–203, 2016.

[9] F. Rossi, R. Zhang, Y. Hindy, and M. Pavone, "Routing autonomous vehicles in congested transportation networks: structural properties and coordination algorithms," *Autonomous Robots*, vol. 42, no. 7, pp. 1427–1442, 2018. [Online]. Available: https://doi.org/10.1007/s10514-018-9750-5

[10] B. Turan, R. Pedarsani, and M. Alizadeh, "Dynamic pricing and fleet management for electric autonomous mobility on demand systems," *Transportation Research Part C: Emerging Technologies*, vol. 121, pp. 1–14, 2020. [Online]. Available: http://arxiv.org/abs/1909.06962

[11] R. Iglesias, F. Rossi, K. Wang, D. Hallac, J. Leskovec, and M. Pavone, "Data-driven model predictive control of autonomous mobility-on-demand systems," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 6019–6025, 2018.

[12] M. Cocca, D. Giordano, M. Mellia, and L. Vassio, "Free Floating Electric Car Sharing: A Data Driven Approach for System Design," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4691–4703, 2019.

[13] M. Gueriau and I. Dusparic, "SAMoD: Shared Autonomous Mobility-on-Demand using Decentralized Reinforcement Learning," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, vol. 2018-Novem, 2018, pp. 1558–1563.

[14] R. M. Swaszek and C. G. Cassandras, "Load Balancing in Mobility-on-Demand Systems: Reallocation Via Parametric Control Using Concurrent Estimation," *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019*, pp. 2148–2153, 2019.

[15] F. He, Y. Yin, and J. Zhou, "Integrated pricing of roads and electricity enabled by wireless power transfer," *Transportation Research Part C: Emerging Technologies*, vol. 34, pp. 1–15, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.trc.2013.05.005

[16] F. He, Y. Yin, J. Wang, and Y. Yang, "Sustainability SI: Optimal Prices of Electricity at Public Charging Stations for Plug-in Electric Vehicles," *Networks and Spatial Economics*, vol. 16, no. 1, pp. 131–154, 2016.

[17] W. Wei, L. Wu, J. Wang, and S. Mei, "Network equilibrium of coupled transportation and power distribution systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6764–6779, nov 2018. [Online]. Available: https://ieeexplore.ieee.org/document/7967870/

[18] W. Lee, R. Schober, and V. W. Wong, "An Analysis of Price Competition in Heterogeneous Electric Vehicle Charging Stations," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3990–4002, jul 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8385151/

[19] W. Yuan, J. Huang, and Y. J. Zhang, "Competitive Charging Station Pricing for Plug-In Electric Vehicles," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 627–639, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7352372/

[20] A. Ghosh and V. Aggarwal, "Menu-Based Pricing for Charging of Electric Vehicles with Vehicle-to-Grid Service," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 268–10 280, nov 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8437164/

[21] A. Moradipari and M. Alizadeh, "Pricing and Routing Mechanisms for Differentiated Services in an Electric Vehicle Public Charging Station

Network," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1489–1499, 2020.

[22] Z. Moghaddam, I. Ahmad, D. Habibi, and M. A. Masoum, "A coordinated dynamic pricing model for electric vehicle charging stations," *IEEE Transactions on Transportation Electrification*, vol. 5, no. 1, pp. 226–238, mar 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8632961/

[23] Z. Zhang, P. Han, J. Wang, Y. Li, and Y. Han, "Agent-based modeling and simulation for the coordinated pricing strategy of the electric vehicle battery switching station," in *Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015*, 2015, pp. 5521–5527.

[24] Z. Ding, Y. Lu, K. Lai, M. Yang, and W. J. Lee, "Optimal coordinated operation scheduling for electric vehicle aggregator and charging stations in an integrated electricity-transportation system," *International Journal of Electrical Power and Energy Systems*, vol. 121, p. 104060, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142061519335616

[25] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, and G. Strbac, "Deep Reinforcement Learning for Strategic Bidding in Electricity Markets," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1343–1355, mar 2020.

[26] M. Salazar, N. Lanzetti, F. Rossi, M. Schiffer, and M. Pavone, "Intermodal Autonomous Mobility-on-Demand," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3946–3960, 2020.

[27] A. M.-o.-d. Systems, "Autonomes Fahren," *Autonomes Fahren*, 2015.

[28] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *35th International Conference on Machine Learning, ICML 2018*, vol. 5, pp. 2976–2989, 2018.

[29] F. Rossi, R. Iglesias, M. Alizadeh, and M. Pavone, "On the Interaction between Autonomous Mobility-on-Demand Systems and the Power Network: Models and Coordination Algorithms," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 1, pp. 384–397, 2020.

[30] S. Kasina and B. F. Hobbs, "The value of cooperation in interregional transmission planning: A noncooperative equilibrium model approach," *European Journal of Operational Research*, vol. 285, no. 2, pp. 740–752, 2020. [Online]. Available: https://doi.org/10.1016/j.ejor.2020.02.018

[31] J. von Neumann and O. Morgenstern, *Theory of games and economic behavior*. Princeton university press, 2007.

[32] M. Simaan and J. B. Cruz, "On the Stackelberg strategy in nonzero-sum games," *Journal of Optimization Theory and Applications*, vol. 11, no. 5, pp. 533–555, 1973.

[33] A. J. Conejo, L. Baringo, S. Jalal Kazempour, and A. S. Siddiqui, *Investment in electricity generation and transmission: Decision making under uncertainty*. Departments of Integrated Systems Engineering, Electrical and Computer Engineering, The Ohio State University, Columbus, OH, United States: Springer International Publishing, 2016. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85016736911&doi=10.1007%2F978-3-319-29501-5&partnerID=40&md5=3f972944e328868c70ae3603cf5db8a4

[34] L. Kraemer and B. Banerjee, "Multi-agent reinforcement learning as a rehearsal for decentralized planning," *Neurocomputing*, vol. 190, pp. 82–94, 2016.

[35] R. Bray, "Markov Decision Processes with Exogenous Variables," *SSRN Electronic Journal*, no. June, 2018.

[36] E. Samadi, A. Badri, and R. Ebrahimpour, "Decentralized multi-agent based energy management of microgrid using reinforcement learning," *International Journal of Electrical Power and Energy Systems*, vol. 122, no. May, p. 106211, 2020. [Online]. Available: https://doi.org/10.1016/j.ijepes.2020.106211

[37] Y. Ye, M. Xiao, and M. Skoglund, "Mobility-Aware Content Preference Learning in Decentralized Caching Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 62–73, 2020.

[38] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-Free Real-Time Autonomous Control for a Residential Multi-Energy System Using Deep Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3068–3082, jul 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9016168/

[39] B. Wang, Y. Li, W. Ming, and S. Wang, "Deep Reinforcement Learning Method for Demand Response Management of Interruptible Load," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3146–3155, 2020.

[40] P. Weber, "User-Generated street Maps," *October*, vol. 7, no. 4, pp. 12–18, 2008.

[41] H. Zhang, S. J. Moura, Z. Hu, W. Qi, and Y. Song, "A Second-Order Cone Programming Model for Planning PEV Fast-Charging Stations," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 2763–2777, 2018.

[42] Y. Liang, C. Guo, Z. Ding, and H. Hua, "Agent-Based Modeling in Electricity Market Using Deep Deterministic Policy Gradient Algorithm," *IEEE Transactions on Power Systems*, vol. 35, no. 6, pp. 4180–4192, nov 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9106862/

[43] M. Powers, M. Shubik, M. R. Powers, and M. Shubik, "The Value of Government and the Efficiency of Noncooperative Equilibrium The Value of Government and the Efficiency of Noncooperative Equilibrium," 2014.

[44] W. K. Hastings, "Monte carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[45] V. François-Lavet, R. Fonteneau, and D. Ernst, "How to Discount Deep Reinforcement Learning: Towards New Dynamic Strategies," pp. 1–9, 2015. [Online]. Available: http://arxiv.org/abs/1512.02011

[46] M. Rayati, M. Toulabi, and A. M. Ranjbar, "Optimal Generalized Bayesian Nash Equilibrium of Frequency-Constrained Electricity Market in the Presence of Renewable Energy Sources," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 1, pp. 136–144, 2020.

[47] J. Mei, X. Xia, and M. Song, "An autonomous hierarchical control for improving indoor comfort and energy efficiency of a direct expansion air conditioning system," *Applied Energy*, vol. 221, no. April, pp. 450–463, 2018. [Online]. Available: https://doi.org/10.1016/j.apenergy.2018.03.162

[48] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–14, 2017.

Ying Lu (Student Member, IEEE) received the B.S. and M.s. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2016 and 2020, respectively. She is currently pursuing the Ph.D. degree at the Department of Electrical Engineering, Technical University of Denmark.

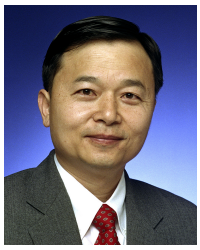
Her current research interests include optimization, reinforcement learning, with applications to power systems and electric transportation systems.

Yanchang Liang (Student Member, IEEE) is a Marie Curie Early Stage Researcher and a PhD student in the School of Engineering at the University of Warwick. Before that, he received the M.S. and B.S. degrees from the School of Electrical Engineering at North China Electric Power University in 2021 and 2018 respectively.

His current research interests include control, optimization, reinforcement learning, with applications to power systems and electric transportation systems.

Zhaohao Ding (Senior member, IEEE) received the B.S. degree in electrical engineering and the B.A. degree in finance both from Shandong University, Jinan, China, in 2010, and the Ph.D. degree in electrical engineering from the University of Texas at Arlington, Arlington, TX, USA, in 2015.

He is currently an Associate Professor with North China Electric Power University, Beijing, China. His research interests include power system planning and operation, power market, distributed resource management, and electric transportation system. He received IEEE IAS Outstanding Young Member Service Award in 2020. He is an Editor of IEEE Transactions on Smart Grid and IEEE Transactions on Industry Applications.



Wei-Jen Lee (Fellow, IEEE) received the B.S. and M.S. degrees from National Taiwan University, Taipei, Taiwan, in 1978 and 1980, respectively, and the Ph.D. degree from The University of Texas at Arlington, Arlington, TX, USA, in 1985, all in electrical engineering. In 1986, he joined the University of Texas at Arlington, where he is currently a Professor with the Department of Electrical Engineering and the Director of the Energy Systems Research Center. He has been involved in research on power flow, transient and dynamic stability, voltage stability, short circuit, relay coordination, power quality analysis, renewable energy, and deregulation for utility companies. Dr. Lee is a registered Professional Engineer in the State of Texas.

Qiuwei Wu (Senior member, IEEE) obtained obtained the PhD degree in Power System Engineering from Nanyang Technological University, Singapore, in 2009. He was a senior RD engineer with VESTAS Technology RD Singapore Pte Ltd from Mar. 2008 to Oct. 2009. He is an Associate Professor at Department of Electrical Engineering, Technical University of Denmark (DTU) since Nov. 2009. He was a visiting scholar at Department of Industrial Engineering Operations Research (IEOR), University of California, Berkeley, from Feb. 2012 to May 2012 funded by Danish Agency for Science, Technology and Innovation (DASTI), Denmark. He was a visiting scholar at School of Engineering and Applied Sciences, Harvard University from Nov. 2017 to Oct. 2018.

His research area is power system operation and control with high renewables, including wind power modelling and control, active distribution networks, and integrated energy systems. He is an Editor of IEEE Transactions on Smart Grid and IEEE Power Engineering Letters. He is also an Associate Editor of International Journal of Electrical Power and Energy Systems, Journal of Modern Power Systems and Clean Energy, IET Renewable Power Generation, and IET Generation, Transmission Distribution.

Tao Ding (Senior member, IEEE) received the B.S.E.E. and M.S.E.E. degrees from Southeast University, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2015. During 2013 and 2014, he was a Visiting Scholar in the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA. He is currently an Associate Professor in the State Key Laboratory of Electrical Insulation and Power Equipment, the School of Electrical Engineering, Xi'an Jiaotong University. His current research interests include electricity markets, power system economics and optimization methods, and power system planning and reliability evaluation. He has published more than 60 technical papers and authored by "Springer Theses" recognizing outstanding Ph.D. research around the world and across the physical sciences—*Power System Operation with Large Scale Stochastic Wind Power Integration*. He received the excellent master and doctoral dissertation from Southeast University and Tsinghua University, respectively, and Outstanding Graduate Award of Beijing City. Dr. Ding is an Editor of IEEE Transactions on Power Systems, IET Generation, Transmission & Distribution, CSEE Journal of Power and Energy Systems.