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: DOI10.1109/TSG.2022.3171656, IEEE Transactions on Smart Grid

A Novel Data-Driven Method for Behind-the-Meter Solar Generation Disaggregation with Cross-Iteration Refinement

Keda Pan, Zhaohua Chen, Chun Sing Lai, *Senior Member, IEEE*, Changhong Xie, Dongxiao Wang, Zhuoli Zhao, *Member, IEEE*, Xiaomei Wu, *Member, IEEE*, Ning Tong, Loi Lei Lai, *Life Fellow, IEEE*, and Nikos D. Hatziargyriou, *Life Fellow, IEEE*

***Abstract*—Photovoltaic (PV) generation is increasing in distribution systems following policies and incentives to promote zero-carbon emission societies. Most residential PV systems are installed behind-the-meter (BTM). Due to single meter deployment that measures the net load only, this PV generation is invisible to distribution system operators causing a negative impact on the distribution system planning and local supply and demand balance. This paper proposes a novel data-driven BTM PV generation disaggregation method using only net load and weather data, without relying on other PV proxies and PV panels' physical models. Long Short-Term Memory (LSTM) is employed to build a generation difference fitted model (GDFM) and a consumption difference fitted model (CDFM) derived from weather data. Both difference fitted models are refined by a cross-iteration with mutual output. Finally, considering the photoelectric conversion properties, the disaggregated generation results are acquired by the refined GDFM of changing input. The proposed method has been tested with actual smart meter data of Austin, Texas and proves to increase the disaggregated accuracy as compared to current state-of-the-art methods. The proposed method is also applicable to disaggregate BTM PV systems of different manufacturing processes and types.**

***Index Terms*—Data-Driven, behind-the-meter, photovoltaic generation disaggregation, machine learning.**

NOMENCLATURE

Abbreviations

BTM	Behind-the-meter
CDFM	Consumption difference fitted model
CV	Coefficient of variation
DHI	Direct normal irradiance
DNI	Diffuse horizontal irradiance
DPVS	Distributed photovoltaic systems
GDFM	Generation difference fitted model
GDM	Generation disaggregation model
GHG	Greenhouse gas
GHI	Global horizontal irradiation
KNN	K-nearest neighbor
LCOE	Levelized cost of energy
LSTM	Long short-term memory
MAPE	Mean absolute percentage error
MSE	Mean squared error
NSRDB	National Solar Radiation Database
PMU	Phasor measurement unit
PV	Photovoltaic
SAM	System advisor model

Parameters

$\mathbf{D}$	Array of date set
$\lvert\mathbf{D}\rvert$	Number of elements in $\mathbf{D}$
iteration	Number of cross-iterations
l	Time step lags
$\mathbf{M}^{d\times n}$	Matrix of date index of d days with n most similar electricity consumption behavior
$\mathbf{M}_i^{d\times n}$	Array of all the elements of row i of matrix $\mathbf{M}^{d\times n}$
$\mathbf{L}^{d\times n}$	Matrix of date index of d days with n most similar solar radiation

This work is sponsored by the National Natural Science Foundation of China (51907031); Guangdong Basic and Applied Basic Research Foundation (Guangdong-Guangxi Joint Foundation) 2021A1515410009; Department of Finance and Education of Guangdong Province 2016 [202]: Key Discipline Construction Program, China; the Education Department of Guangdong Province: New and Integrated Energy System Theory and Technology Research Group [Project Number 2016KCXTD022]; Brunel University London BRIEF Funding. *(Corresponding authors: Chun Sing Lai and Loi Lei Lai)*

K. Pan and L. L. Lai are with the Department of Control Engineering, School of Automation, Guangdong University of Technology, Guangzhou, 510006, China and also with the Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou, 510006, China (e-mail: 1111904017@mail2.gdut.edu.cn; l.l.lai@gdut.edu.cn).

Z. Chen, C. Xie, Z. Zhao, X. Wu and N. Tong are with the Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou, 510006, China (e-mail: 2111904198@mail2.gdut.edu.cn; 2111904168@mail2.gdut.edu.cn; zhuoli.zhao@gdut.edu.cn; epxm_wu@gdut.edu.cn; tongning@gdut.edu.cn).

C. S. Lai is with Brunel Interdisciplinary Power Systems (BIPS) Research Centre, Department of Electronic and Electrical Engineering, Brunel University London, London, UB8 3PH, UK and also with the Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou, 510006, China (e-mail: chunsing.lai@brunel.ac.uk).

D. Wang is with System Design and Engineering Department, Australia Energy Market Operator, Melbourne, 3000, Australia (e-mail: dongxiaouon@gmail.com).

N. D. Hatziargyriou is with the School of Electrical and Computer Engineering, National Technical University of Athens, 15773 Athens, Greece (e-mail: nh@power.ece.ntua.gr).

Copyright © 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

$\mathbf{L}_i^{d\times n}$	Array of all the elements of row i of matrix $\mathbf{L}^{d\times n}$
$\mathbf{T}$	Array of hour set
$\lvert\mathbf{T}\rvert$	Number of elements in $\mathbf{T}$

Variables

$\mathbf{C}_{\mathbf{D},\mathbf{T}}$	Matrix of consumption data of hour set $\mathbf{T}$, date set $\mathbf{D}$ in kW
$\hat{\mathbf{C}}_{i,\mathbf{T}-l}$	Array of estimated consumption of hour set $\mathbf{T}$, date i for l time step lags in kW
$\mathbf{DHI}$	Matrix of direct normal irradiance in $\mathrm{kW}\cdot\mathrm{m}^{-2}$
$\mathbf{DNI}$	Matrix of diffuse horizontal irradiance in $\mathrm{kW}\cdot\mathrm{m}^{-2}$
$\mathbf{G}_{\mathbf{D},\mathbf{T}}$	Matrix of PV generation data of hour set $\mathbf{T}$, date set $\mathbf{D}$ in kW
$\hat{\mathbf{G}}_{i,\mathbf{T}}$	Array of estimated generation of hour set $\mathbf{T}$, date i in kW
$\mathbf{N}_{\mathbf{D},\mathbf{T}}$	Matrix of net load data of hour set $\mathbf{T}$, date set $\mathbf{D}$ in kW
R^2	Coefficient of determination
$\boldsymbol{\tau}_{\mathbf{D},\mathbf{T}}$	Matrix of temperature data of hour set $\mathbf{T}$, date set $\mathbf{D}$ in kW
$\mathbf{X}$	Matrix concatenation of PV-related data
$\mathbf{Y}$	Matrix concatenation of load-related data
$\mathbf{Z}$	Matrix concatenation of $\mathbf{X}$ and estimated lagged historical consumption
ΔC	Consumption difference in kW
$\Delta\mathbf{C}$	Matrix of consumption difference in kW
$\Delta\mathbf{C}_{ij,\mathbf{T}}$	Array of consumption difference of hour set $\mathbf{T}$ between date i and date j in kW
$\Delta\hat{\mathbf{C}}_{ij,\mathbf{T}}$	Array of estimated consumption difference of hour set $\mathbf{T}$ between date i and date j in kW
ΔG	Generation difference in kW
$\Delta\mathbf{G}$	Matrix of generation difference in kW
$\Delta\mathbf{G}_{ij,\mathbf{T}}$	Array of generation difference of hour set $\mathbf{T}$ between date i and date j in kW
$\Delta\hat{\mathbf{G}}_{ij,\mathbf{T}}$	Array of estimated generation difference of hour set $\mathbf{T}$ between date i and date j in kW
ΔN	Net load difference in kW
$\Delta\mathbf{N}_{ij,\mathbf{T}}$	Array of net load difference of hour set $\mathbf{T}$ between date i and date j in kW

Functions

$f_C(\cdot)$	Function of consumption difference fitted model
$f_G(\cdot)$	Function of generation difference fitted model
$L(\cdot)$	Loss function of the neural network

I. Introduction

EXTREME weather caused by excessive greenhouse gas emissions has urged the adoption of measures to combat climate change. In the past decade over 100 countries have set or are considering net-zero Greenhouse Gas (GHG) emissions or neutrality targets and have accelerated the penetration of renewable energy systems in their energy mix [1]. Photovoltaic (PV) technology is one of the most promising sources of green and sustainable energy with low levelized cost of energy (LCOE) [2]. As utility-scale PV installations have slowed down in 2020, distributed PV continues to increase, driven by steady interest in self-consumption, the benefits of net metering and technical advantages in improving reliability and reducing network losses [3], [4]. Most distributed PVs are installed behind-the-meter (BTM), with only net load visible to utility operators due to single meter deployment restrictions [5] and power information privacy issues of customers [6]. The unavailability of separate generation and consumption can cause several problems to distribution system operators, such as inaccurate load forecasting and baseline load estimation [7], [8], suboptimal distribution systems planning [9], [10] and incorrect hosting capacity analysis [11]. To avoid these problems, several BTM net load disaggregation methods have been proposed.

The existing BTM net load disaggregation methods can be divided into data-driven and model based methods. Data-driven methods can be further divided into energy-proxy-dependent and energy-proxy-independent ones.

Energy-proxy-dependent data-driven methods: Because of the invisibility of BTM PV, some researchers utilize the generation and consumption characteristics of the representative energy agents to explain the target BTM customers, indirectly obtaining the disaggregation results. By inputting the target capacity into the fitted function trained by installation capacity and PV generation of publicly available PV sites proxy, the BTM PV is disaggregated in [12], [13]. With the help of multiple linear models regression strategy, [14] the consumption and generation, represented by phasor measurement unit (PMU) recorded reactive power and output of nearby PV sites proxy are estimated, by minimizing the errors between the estimated and monitored net load of feeder level. In [15], the installation capacity of BTM PV was firstly inferred by a support vector regression model trained by the output characteristics of a nearby unit capacity PV site proxy. Then the generation of BTM PV is approximated by its capacity multiplied by the output power of the standard PV sites. Instead of utilizing PV sites proxy, the "consumption" proxy was introduced in [16] and [17]. By assuming that the consumption behavior of a consumer is a mixture of latent behaviors, this method infers the full-day consumption of the BTM customer through a linear function fitted by the nighttime load of mixture proxy customers. In [18], [5], by utilizing contextually supervised source separation, a PV disaggregation model was designed solving a simplified linear problem with scale generation and consumption data of proxy customers as additional input. Based on the high correlation of load among neighboring customers, when approximated by environmental and social factors, and the high correlation of neighboring PV sites, when approximated by meteorological factors, [19] the residential BTM PV generation is estimated employing spatiotemporal graph with neighboring net load as input which

can be considered as specific proxies.

Disaggregation methods based on energy-proxy-dependent sites have an easy-to-understand model basis, but they require a consistent behavior of consumption and generation characteristics between energy proxies and target users. In practice, there are collection time requirements of data after setting the proxies, making the disaggregation methods not immediately available.

Energy-proxy-independent data-driven methods: Some data-driven methods do not use a proxy-based approach in order to avoid the defective disaggregation results caused by heterogeneous data distribution. In [20], a small number of representative solar sites was selected according to a four-month PV generation data of all sites using a dimension reduction method. The aggregated generation estimation is obtained by a supervised learning method built by the representative data and external variables, such as weather and locational information. In the case of known historical BTM PV generation data, an exemplar library composed of fully observable customers is established in [21]. The disaggregation of PV is obtained by the combination of typical customer generation patterns optimized by game-theoretic approaches. With partial BTM generation data being labeled, the BTM net load is disaggregated in a dictionary learning manner in [22]. In [23], when the approximate installation time of the PV sites is known, the BTM consumption can be obtained from the load data before the PV installation determined by the time stamp comparison of relevant factors such as weather, day type, etc., and further combined with net load estimate of BTM PV generation.

The above-mentioned methods adopt their own historical separate metering data for disaggregation, which to a certain extent avoids the problem of data heterogeneity brought by setting proxies, but in reality, their implementation suffers from hindrances. These are caused by factors such as the large number of small-distributed PV systems (DPVS) that use net metering or changes in consumption behaviors after the installation of DPVS. Hence, the study reported in [24] has developed an unsupervised disaggregation method that does not rely on separate historical consumption and generation data. By taking demand and irradiance-generation transposition parameters as the unknown variables and minimizing the estimated net load and actual net load as the goal, this method disaggregates the BTM PV using multiple linear programming.

Model based methods: In [25], the BTM PV disaggregation results were obtained according to the clear sky generation modified by the universal weather-solar effect. This considers that under the same meteorological conditions, reduction of the maximum clear sky irradiation is the same. The construction of the disaggregation model requires a search for the installation angle of PV equipment, and the minimum power consumption floor for users based on insufficient assumptions. In [7], depending on the PV system geometry, the estimated output distribution was first obtained by combining different equivalent parameters of capacity, tilt angle and azimuth. Then, the disaggregated PV results inferred according to the optimal equivalent parameter combination is determined, when the estimated PV generation has the least correlation with the estimated residual. In [26], the estimated generation and residential consumption were calculated by customized PV system physical model and hidden Markov regression model, respectively. The estimation results of the generation and consumption are combined with net load to feed the mutual model for iterative updating of the model parameters. Further improvements of [26] were presented in [27], where hidden Markov model is changed to mixed hidden Markov model to enable modelling the general load consumption behavior present in population-level, while acknowledging the differences of individual loads. This method can also estimate the parameters of multiple strings of solar panels and different types of sites.

Model-based algorithms are less data-dependent, which is advantageous for implementation, but the assumptions of the physical model must be realistic. However, different PV panel manufacturing processes (mono-c-Si, multi-c-Si and thin film) and different types of PV system (monofacial, bifacial, tracking or hybrid) can both cause the physical model to vary. At the same time, it is difficult to know the specific PV system information of the customer in advance, because it involves personal privacy issues and there may be misinformation, self-installation, and aging of the DPVS. The advantages and disadvantages of the different categories of BTM net load disaggregation methods are summarized in Table I

TABLE I
THE ADVANTAGES AND DISADVANTAGES OF EXISTING BTM NET LOAD DISAGGREGATION METHODS

Category	Advantages	Disadvantages
Energy proxy-dependent data-driven methods	Knowledge of physical models not necessary; easy to understand and model; supervised learning feasible; strong generalization ability	Setting of the proxies needs to be representative; time costs of the collection data of proxies
Energy proxy-independent data-driven methods		Most methods require sub-metering of data
Model based method	Less data-dependent; changes in behaviors of customers have less impact on disaggregation results	Prior knowledge of physical model; physical model assumptions are prone to inconsistencies with target customers; highly influenced by physical changes in the PV system

From the algorithmic point of view, data-driven methods of [14], [16], [18], and [24] apply a simplified linear model in the calculation process, which is difficult to fully reflect the relationship between the variables. In model-based methods of [7], [26], and [27], the determination of the optimal solution and the iterative update of the methods are based on the model error decision of the constructed regression algorithm. The iterative process of the methods lacks feedback on model updates, which can lead to long computation times and falling into local

optimum. For ease of comparison, the model properties of different methods are summarized in Table II.

TABLE II
THE MODEL PROPERTIES OF EXISTING BTM NET LOAD DISAGGREGATION METHODS

Model properties	Methods													
	[5], [18]	[7]	[12]	[13], [15]	[14]	[16], [17]	[19]	[20]	[21], [22]	[23]	[24]	[25]	[26], [27]	Proposed Method
Energy-proxy-dependent data-driven method	√	×	√	√	√	√	√	×	×	×	×	×	×	×
Energy-proxy-independent data-driven method	×	×	×	×	×	×	×	√	√	√	√	×	×	√
Model based method	×	√	×	×	×	×	×	×	×	×	×	√	√	×
PV sites proxy	√	×	√	√	√	×	√	×	×	×	×	×	×	×
"Consumption" proxy	√	×	×	×	×	√	√	×	×	×	×	×	×	×
Need historical separate metering data	×	×	×	×	×	×	×	√	√	√	×	×	×	×
Need meteorological data	√	√	×	√	×	√	×	×	×	√	√	√	√	√
Need PV physical model parameters	×	√	√	√	×	×	×	×	×	×	×	√	×	×
Need the installation information	×	×	×	×	×	×	×	√	×	×	√	×	×	×
Need searching for the installation parameters of PV system geometry	×	√	×	×	×	×	×	×	×	×	×	√	√	×
Simplified linear relationship model	√	×	×	×	√	√	×	×	×	√	√	×	×	×
Machining learning based model	×	√	×	√	×	×	√	√	√	×	×	×	√	√
Application of BTM disaggregation to a hybrid of different PV panel manufacturing processes and different types of PV system scenarios without prior knowledge of specific situation	×	×	×	×	×	×	×	×	×	×	×	×	×	√

Data-driven methods have the advantage over model-based methods of not being bound by the assumptions of a PV physical model. As seen from Table II, all data-driven BTM PV disaggregation methods require either proxies setup or historical separate metering data. However, in BTM PV installations, historical separate metering data is frequently unknown, thus it is difficult to employ supervised learning for acquiring disaggregation results without setting up an energy proxy. A proxy setup has the same quandary as in the model based methods, namely whether the energy proxies or the physical model assumptions can effectively represent the target to be disaggregated. To avoid these difficulties, this paper proposes a novel data-driven BTM PV generation disaggregation approach using only net load and weather data, without relying on PV proxies and physical models assumptions of PV panels.

More specifically, for a prosumer, characterized by the same approximate consumption behavior under similar weather conditions and day types, and the same approximate PV generation under the same meteorological conditions, the deep learning based initial consumption difference fitted model (CDFM) and generation difference fitted model (GDFM) are constructed. The pairwise net load difference is obtained as output, while their exogenous variables data is used as input. Since it is difficult to find identical days of consumption behavior or meteorological conditions, the offset in consumption and generation of the net load data is compensated according to the exogenous variables. The outputs of the CDFM are used to correct the supervised target of the GDFM and vice versa. This process will continue to update mutual fitted models in a cross-iteration ways until the disaggregation results no longer change significantly. After iterations, a "pure" GDFM is obtained. The model is converted to a generation disaggregation model (GDM) by changing the inputs to obtain estimated BTM PV generation results based on the characteristic that the PV output is zero, when the irradiation is zero.

The original contributions of this paper are as follows:

1) A novel BTM GDM is obtained from the GDFM by transfer from net load difference regression to generation disaggregation changing the model inputs. The proposed data-driven method requires only net load and nearby weather data, avoiding the problem of performance inconsistency between the assumed physical PV model and the target BTM one, typically found in model-based disaggregation methods. Also, the proposed method does not need an energy-proxy nor historical separate metering data. The exogenous variables are directly passed to the disaggregation target through the LSTM network, avoiding transposition errors.

2) In the proposed method, supervised targets of GDFM and CDFM are extracted by only net load data. . GDFM and CDFM are trained with supervised learning by pairwise net load differences according to the weather and day type data. Both fitted models are refined by cross-iteration to enhance their practical interpretability. In the iterative process, the estimated historical consumption is calculated from the estimated PV generation obtained from the GDM and used as time lags load sequences features for input to obtain a CDFM with stronger interpretation and higher accuracy.

3) Based on a data-driven approach and without prior knowledge of the number of PV sites, installation configurations and physical models; the proposed method can disaggregate the net load with DPVS installation of different

panel manufacturing processes (mono-c-Si, multi-c-Si and thin film), different types (monofacial, bifacial, not tracking and tracking) or hybrid types, which can be conveniently applied for various scenarios.

The paper is organized as follows: Section II presents the disaggregation methodology. Section III presents the case study results and discusses the model performance. Section IV provides conclusions and future work.

II. Methodology

In a BTM system, PV generation data $\mathbf{G_{D,T}}$ (kW) is invisible, while only net load dataset $\mathbf{N_{D,T}}$ (kW) is known. This equals the consumption $\mathbf{C_{D,T}}$ (kW) minus the PV generation $\mathbf{G_{D,T}}$.

$$\mathbf{N_{D,T}} = \mathbf{C_{D,T}} - \mathbf{G_{D,T}} \tag{1}$$

The subscripts $\mathbf{D}$ and $\mathbf{T}$ represent the set of date and set of hour, respectively. $\mathbf{C_{D,T}}$ is unknown and not negligible, therefore, it is impossible to directly build a neural network model without training targets, i.e., the relationship between PV-related input and $\mathbf{G_{D,T}}$.

When the consumption difference between two days $\Delta\mathbf{C}$ (kW) is known (one of the key problems to be solved in this research), the generation difference $\Delta\mathbf{G}$ (kW) between the two days can be calculated as in (2):

$$\Delta\mathbf{G}_{ij,\mathbf{T}} = \Delta\mathbf{C}_{ij,\mathbf{T}} - \Delta\mathbf{N}_{ij,\mathbf{T}}, \quad i \in \mathbf{D}, j \in \mathbf{D}, i \neq j \tag{2}$$

$$\begin{cases} \Delta\mathbf{G}_{ij,\mathbf{T}} = \mathbf{G}_{i,\mathbf{T}} - \mathbf{G}_{j,\mathbf{T}} \\ \Delta\mathbf{C}_{ij,\mathbf{T}} = \mathbf{C}_{i,\mathbf{T}} - \mathbf{C}_{j,\mathbf{T}} \\ \Delta\mathbf{N}_{ij,\mathbf{T}} = \mathbf{N}_{i,\mathbf{T}} - \mathbf{N}_{j,\mathbf{T}} \end{cases} \tag{3}$$

A GDFM $f_G(\cdot)$ can be built by neural networks with input the irradiance data and output the estimated difference of PV generation $\Delta\hat{\mathbf{G}}_{ij}$ (kW):

$$\Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} = f_G(\mathbf{X}_{i,\mathbf{T}}, \mathbf{X}_{j,\mathbf{T}}) \tag{4}$$

$$\begin{cases} \mathbf{X}_{i,\mathbf{T}} = [\mathbf{DNI}_{i,\mathbf{T}}, \mathbf{DHI}_{i,\mathbf{T}}] \\ \mathbf{X}_{j,\mathbf{T}} = [\mathbf{DNI}_{j,\mathbf{T}}, \mathbf{DHI}_{j,\mathbf{T}}] \end{cases} \tag{5}$$

where $\mathbf{DNI}$ and $\mathbf{DHI}$ $\left(\mathrm{kW \cdot m^{-2}}\right)$ represent the direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI) at the corresponding time, respectively.

The loss function $L(\cdot)$ of the neural network can be calculated from $\Delta\mathbf{G}_{ij,\mathbf{T}}$ and $\Delta\hat{\mathbf{G}}_{ij,\mathbf{T}}$ as shown in (6).

$$L(\Delta\hat{\mathbf{G}}_{ij,\mathbf{T}}, \Delta\mathbf{G}_{ij,\mathbf{T}}) = \sum_{i,j \in \mathbf{D}} \left(\Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} - \Delta\mathbf{G}_{ij,\mathbf{T}}\right)^2 \tag{6}$$

Due to the photoelectric conversion properties of PV panels, the PV generation is zero, when DNI and DHI are zero [28]. By changing $\mathbf{X}_{j,\mathbf{T}}$ to $\mathbf{0}$, the estimated PV generation $\hat{\mathbf{G}}_{i,\mathbf{T}}$ can be obtained through GDM converted by GDFM from (4) as follows:

$$\begin{cases} \hat{\mathbf{G}}_{i,\mathbf{T}} = \Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} = f_G(\mathbf{X}_{i,\mathbf{T}}, \mathbf{0}) \\ \hat{\mathbf{G}}_{j,\mathbf{T}} = \mathbf{0}, \ s.t. \ \mathbf{X}_{j,\mathbf{T}} = \mathbf{0} \end{cases} \tag{7}$$

It is worth mentioning that the input features of $f_G(\cdot)$ are not limited to $\mathbf{X}_{i,\mathbf{T}}$ and $\mathbf{X}_{j,\mathbf{T}}$; it can be extended to other PV-related inputs.

Regarding the unknown $\Delta\mathbf{C}$ in (2), there are several methods to obtain the solutions, as described in the following subsections.

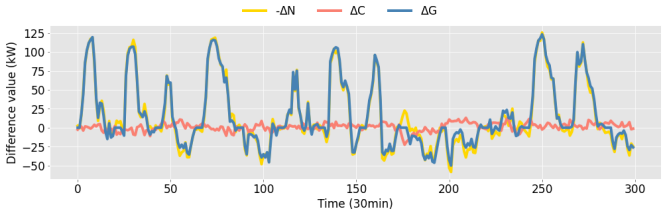
A. Method A

For the unknown $\Delta\mathbf{C}$, we first consider a net load difference of two days with similar electricity consumption behaviors. According to the continuity of electricity consumption behaviors of customers, we assume that the net load curves with similar shape at nighttime will also have similar shape during the daytime. In order to eliminate the influence of generation information on the day matching of electricity consumption behavior, the n days with most similar electricity consumption behavior at nighttime among d days are searched using the k-nearest neighbor (KNN) algorithm, and the corresponding date index is recorded in the matrix $\mathbf{M}^{d\times n}$. When training GDFM, the date index of day i is denoted as $\mathbf{M}_i^{d\times n}$, which represents all the elements of row i of matrix $\mathbf{M}^{d\times n}$.

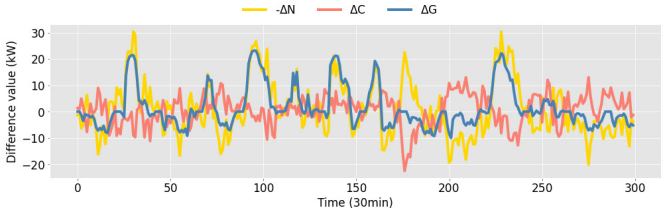
Thus, the GDFM can be modeled according to (8) with the loss function (6):

$$\begin{cases} \Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} = f_G(\mathbf{X}_{i,\mathbf{T}}, \mathbf{X}_{j,\mathbf{T}}) \\ \Delta\mathbf{G}_{ij,\mathbf{T}} = -\Delta\mathbf{N}_{ij,\mathbf{T}} \\ s.t. \ \Delta\mathbf{C}_{ij,\mathbf{T}} = \mathbf{0} \end{cases}, \ i \in \mathbf{D}, j \in \mathbf{M}_i^{d\times n} \tag{8}$$

When the PV penetration is relatively high, this approximation is reasonable, and satisfactory disaggregation results can be obtained. Fig. 1(a) demonstrates $\Delta\mathbf{C}$, which is negligible at high PV penetration.



(a) Scenario of 150% PV penetration



(b) Scenario of 30% PV penetration

Fig. 1. Profiles of ΔN, ΔC and ΔG in Method A.

B. Method B

In low PV penetration scenarios, the influence of $\Delta\mathbf{C}$ is significant, see Fig. 1(b). Thus, a neural network for CDFM is developed, which can refine the GDFM by compensating for $\Delta\mathbf{C}$.

Similar to GDFM, (9) can be obtained by rewriting (2):

$$\Delta\mathbf{C}_{ij,\mathbf{T}} = \Delta\mathbf{G}_{ij,\mathbf{T}} + \Delta\mathbf{N}_{ij,\mathbf{T}}, \quad i \in \mathbf{D}, j \in \mathbf{D}, i \neq j \tag{9}$$

Since the PV generation is highly related to solar radiation, it can be considered that when the irradiation conditions are similar in two days, the PV generation is also similar. By finding the days with similar irradiation, the n most similar PV generation days among d days are searched through KNN according to DHI and DNI, and the corresponding date index is recorded in the matrix $\mathbf{L}^{d\times n}$. When training the CDFM, the date index of day i is denoted as $\mathbf{L}_i^{d\times n}$, which represents all the elements of row i of matrix $\mathbf{L}^{d\times n}$.

To build the CDFM, $\Delta\mathbf{G}$ is offset by obtaining a net load difference of two days with similar PV generation according to the date index of $\mathbf{L}^{d\times n}$. CDFM is modelled as (10), by using neural networks with the loss function (11):

$$\begin{cases} \Delta\hat{\mathbf{C}}_{ij,\mathbf{T}} = f_C(\mathbf{Y}_{i,\mathbf{T}}, \mathbf{Y}_{j,\mathbf{T}}) \\ \Delta\mathbf{C}_{ij,\mathbf{T}} = \Delta\mathbf{N}_{ij,\mathbf{T}} \quad , i \in \mathbf{D}, j \in \mathbf{L}_i^{d\times n} \\ s.t. \Delta\mathbf{G}_{ij,\mathbf{T}} = \mathbf{0} \end{cases} \tag{10}$$

$\mathbf{Y}$ is load-related exogenous data, such as temperature (C°) and date type. It is worth mentioning that there is no fixed standard for the input features of CDFM. Similar to load forecasting, the selection of input features can be done in the same way as existing load forecasting models, except that the CDFM model solves for the regression of consumption differences, rather than the regression of consumption.

The loss function in the neural network of CDFM can be calculated from $\Delta\mathbf{C}_{ij,\mathbf{T}}$ and the estimated consumption difference $\Delta\hat{\mathbf{C}}_{ij,\mathbf{T}}$ (kW):

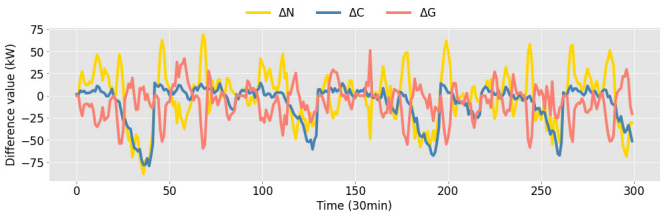
$$L(\Delta\hat{\mathbf{C}}_{ij,\mathbf{T}}, \Delta\mathbf{C}_{ij,\mathbf{T}}) = \sum_{i,j\in\mathbf{D}} \left(\Delta\hat{\mathbf{C}}_{ij,\mathbf{T}} - \Delta\mathbf{C}_{ij,\mathbf{T}}\right)^2 \tag{11}$$

The GDFM can be compensated by CDFM to calculate $\Delta\mathbf{C}$ according to the load-related features with the same date index of $\mathbf{M}^{d\times n}$. Thus, the GDFM can still be modelled as (12) with the loss function (6):

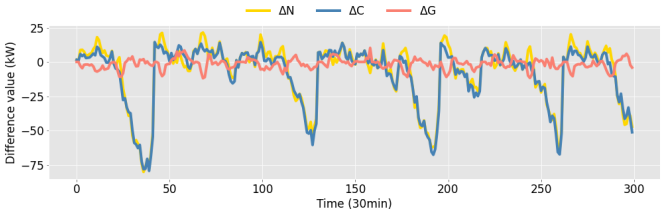
$$\begin{cases} \Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} = f_G(\mathbf{X}_{i,\mathbf{T}}, \mathbf{X}_{j,\mathbf{T}}) \\ \Delta\mathbf{G}_{ij,\mathbf{T}} = -\Delta\mathbf{N}_{ij,\mathbf{T}} + \Delta\hat{\mathbf{C}}_{ij,\mathbf{T}} \ , i \in \mathbf{D}, j \in \mathbf{M}_i^{d\times n} \\ s.t. \ \Delta\hat{\mathbf{C}}_{ij,\mathbf{T}} = f_C(\mathbf{Y}_{i,\mathbf{T}}, \mathbf{Y}_{j,\mathbf{T}}) \end{cases} \tag{12}$$

C. Method C

Similar to GDFM in Method A, the CDFM in Method B of (10) assumes that $\Delta\mathbf{G} = 0$, which is acceptable at low PV penetration, as shown in Fig. 2(b). In the case of high PV penetration, the influence of $\Delta\mathbf{G}$ is important, see Fig. 2(a).



(a) Scenario of 150% PV penetration



(b) Scenario of 30% PV penetration

Fig. 2. Profiles of ΔN, ΔC and ΔG in Method B.

Therefore, after the initial GDFM is obtained from Method A, the output of CDFM can be compensated by GDFM, in turn, in a form similar to (12):

$$\begin{cases} \Delta\hat{\mathbf{C}}_{ij,\mathbf{T}} = f_C(\mathbf{Y}_{i,\mathbf{T}}, \mathbf{Y}_{j,\mathbf{T}}) \\ \Delta\mathbf{C}_{ij,\mathbf{T}} = \Delta\mathbf{N}_{ij,\mathbf{T}} + \Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} \quad , i \in \mathbf{D}, j \in \mathbf{L}_i^{d\times n} \\ s.t. \Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} = f_G(\mathbf{X}_{i,\mathbf{T}}, \mathbf{0}) - f_G(\mathbf{X}_{j,\mathbf{T}}, \mathbf{0}) \end{cases} \tag{13}$$

The estimated PV generation values $\hat{\mathbf{G}}$ of (7) are used directly in (13) instead of the $\Delta\mathbf{G}$ in GDFM. Since, they directly correspond to the final generation target, the final estimated generation results have higher robustness.

It needs to be emphasized that, in Method B, the CDFM can only be constructed from load-related exogenous data (temperature and data type) due to the invisible historical consumption data $\mathbf{C}$, which is difficult to adequately account for the variation in load difference. However, in Method C, since $\hat{\mathbf{G}}$ is first estimated from (7), the estimated consumption $\hat{\mathbf{C}}$ can be obtained from (1) and used for the model building of CDFM in Method C by rewriting (13).

$$\begin{cases} \Delta\hat{\mathbf{C}}_{ij,\mathbf{T}} = f_C(\mathbf{Z}_{i,\mathbf{T}-l}, \mathbf{Z}_{j,\mathbf{T}-l}) \\ \Delta\mathbf{C}_{ij,\mathbf{T}} = \Delta\mathbf{N}_{ij,\mathbf{T}} + \Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} \quad , i \in \mathbf{D}, j \in \mathbf{L}_i^{d\times n} \\ s.t. \Delta\hat{\mathbf{G}}_{ij,\mathbf{T}} = f_G(\mathbf{X}_{i,\mathbf{T}}, \mathbf{0}) - f_G(\mathbf{X}_{j,\mathbf{T}}, \mathbf{0}) \end{cases} \tag{14}$$

$$\begin{cases} \mathbf{Z}_{i,\mathbf{T}}^l = [\mathbf{Y}_{i,\mathbf{T}}, \hat{\mathbf{C}}_{i,\mathbf{T}-1}, \ldots, \hat{\mathbf{C}}_{i,\mathbf{T}-l}] \\ \mathbf{Z}_{j,\mathbf{T}}^l = [\mathbf{Y}_{j,\mathbf{T}}, \hat{\mathbf{C}}_{j,\mathbf{T}-1}, \ldots, \hat{\mathbf{C}}_{j,\mathbf{T}-l}] \end{cases} \tag{15}$$

$$\begin{cases} \hat{\mathbf{C}}_{i,\mathbf{T}-l} = \mathbf{N}_{i,\mathbf{T}-l} + \hat{\mathbf{G}}_{i,\mathbf{T}-l} \\ \hat{\mathbf{C}}_{j,\mathbf{T}-l} = \mathbf{N}_{j,\mathbf{T}-l} + \hat{\mathbf{G}}_{j,\mathbf{T}-l} \end{cases} \tag{16}$$

where l represents the number of time step lag. The CDFM of (14) constructed by additionally considering the continuous time lags sequences of estimated historical consumption $\hat{\mathbf{C}}$ itself can track and learn the temporal relationship of energy consumption difference, has a higher quality of regression of

the estimated consumption difference values compared to CDFM of (13).

A cross-iteration process can be formed by combining (7), (12) and (14). This process can be initialized by (8), that is, the outputs of GDM compensate for the supervised targets of CDFM, and the outputs of CDFM are used as compensation of supervised targets of GDFM at the next stage. It is worth noting that in Method C, due to the presence of cross-iterations, the pairwise net load can be randomly selected for matching and the indices of the most similar consumption behavior days $\mathbf{M}^{d\times n}$ and irradiation days $\mathbf{L}^{d\times n}$ are not necessary. The compensation of supervised targets will gradually refine GDFM and CDFM simplifying the overall modelling process.

The algorithm of Method C is summarized below.

Algorithm 1: Algorithm for solar generation disaggregation with cross-iteration refinement

Input: Net load dataset $\mathbf{N}_{\mathbf{D},\mathbf{T}}$, weather dataset $\mathbf{DNI}_{\mathbf{D},\mathbf{T}}$, $\mathbf{DHI}_{\mathbf{D},\mathbf{T}}$ and $\boldsymbol{\tau}_{\mathbf{D},\mathbf{T}}$.

Output: Estimated PV generation $\hat{\mathbf{G}}_{\mathbf{D},\mathbf{T}}$,

1: Randomly match the net load and generate date indices $\mathbf{M}^{d\times n}$ and $\mathbf{L}^{d\times n}$ for GDFM and CDFM respectively.
2: Train the initial GDFM model according to Equation (8) with $\mathbf{M}^{d\times n}$ to obtain $f_G^{(0)}(\cdot)$.
3: Obtain the estimated PV generation $\hat{\mathbf{G}}_{\mathbf{D},\mathbf{T}}^{(0)}$ from $f_G^{(0)}(\cdot)$ according to Equation (7).
4: **for** $k=1$ to *iteration* **do**
5: Train CDFM according to Equation (13) with $\mathbf{L}^{d\times n}$, obtain $f_C^{(k)}(\cdot)$.
6: Obtain $\Delta\hat{\mathbf{C}}_{\mathbf{D},\mathbf{T}}$ from $f_C^{(k)}(\cdot)$ with $\mathbf{M}^{d\times n}$ in Equation (14).
7: Train GDFM according to Equation (12) with $\mathbf{M}^{d\times n}$, obtain $f_G^{(k)}(\cdot)$.
8: Obtain the estimated PV generation $\hat{\mathbf{G}}_{\mathbf{D},\mathbf{T}}^{(k)}$ from $f_G^{(k)}(\cdot)$ according to Equation (7).
9: **end for**
10: **return** $\hat{\mathbf{G}}_{\mathbf{D},\mathbf{T}}^{(epoch)}$

III. Case Studies

A. Dataset and Experimental Setting

To verify the effectiveness of the proposed method in real-life applications, open source residential net load data in Austin Texas [29] are used. The PV generation was recorded separately. After data filtering and complementing, the data of 24 (originally 25, the consumption of a customer was metered negative numbers) customers are retained for the disaggregation experiments. The metering time of the dataset is from 01/01/2018 to 30/12/2018 with a 30-minute interval and the total nameplate capacity is 102.53kWp. Due to customers' privacy issues, the specific geographic coordinates of the residence are not given. The required meteorological data for the corresponding period are obtained from the National Solar Radiation Database (NSRDB) [30] of Austin, Texas at approximate latitude of 30.25° N and longitude of -97.74° E. To test the disaggregation effect of the proposed approach, this is applied to net loads with DPVS of different manufacturing processes and types; semi-synthesized data are introduced by eliminating the generation part of the data obtained from [29] and adding the customized PV data calculated by System Advisor Model (SAM) [31].

The experiments in case studies are implemented by Python 3.7.7 on a server with NVIDIA Geforce RTX 2080Ti GPU and 64 GB of RAM. Deep learning model is implemented based on the TensorFlow 1.13.1 framework.

GDFM and CDFM are essentially time series regression modelling. To improve the explanatory ability of these models without losing generality, the LSTM is implemented for modeling [32]. The hyperparameters of LSTM for GDFM and CDFM are listed in the Appendix. The mean squared error (MSE) and the coefficient of variation (CV) are used as evaluation metrics of disaggregation accuracy. Considering that small disaggregation errors can cause great fluctuations of mean absolute percentage error (MAPE) when the output is close to zero (e.g., cloudy days or sunrise and sunset), CV is used instead of MAPE [18]. The MSE and CV are defined as:

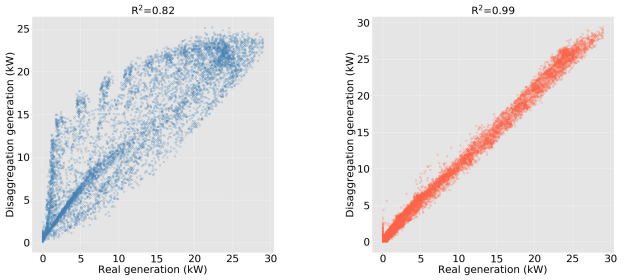
$$MSE=\frac{1}{|\mathbf{D}||\mathbf{T}|}\sum_{d\in\mathbf{D}}\sum_{t\in\mathbf{T}}\left(G_{d,t}-\hat{G}_{d,t}\right)^2 \tag{17}$$

$$CV=\frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}}\left(\frac{\sqrt{\sum_{t\in\mathbf{T}}\left(G_{d,t}-\hat{G}_{d,t}\right)^2}}{\sum_{t\in\mathbf{T}}G_{d,t}}\right) \tag{18}$$

where $|\mathbf{D}|$ and $|\mathbf{T}|$ represent the number of the elements of array $\mathbf{D}$ and array $\mathbf{T}$.

B. Learning Task Migration

The key point of the proposed data-driven disaggregation algorithm is whether GDFM can be converted into GDM by setting the unconcerned irradiation values of the pairwise inputs to zero. To illustrate the feasibility of "migration", we build the GDFM using generation data only and excluding the effect of consumption in net load. Fig. 3 illustrates the scatter plot between the real generation and disaggregation generation obtained by migration. Coefficient of determination R^2 is used to explain the model regression ability.



(a) DNI and DHI as input feature (b) DNI, DHI, hour angle as input feature

Fig. 3. Scatter plot between the real generation and disaggregation generation of different input feature.

The regression performance is shown in Fig. 3 after GDFM is converted to GDM. A comparison between Fig. 3(a) and Fig. 3(b) shows that the R^2 value improves from 0.82 to 0.99 by adding unilateral variables as input variables. This effectively improves the regression accuracy and demonstrates the value of adding PV-related variables to the concerned side, as shown in Equation (7).

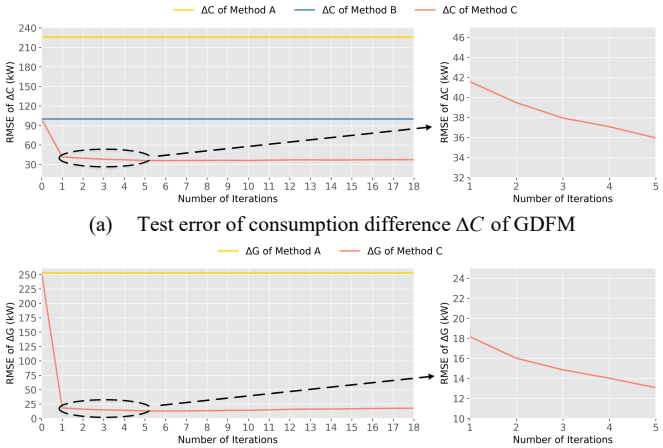
C. *Methods Performance and Comparison*

The core idea behind the improvement of Methods A and B to obtain C is the reasonable elimination of ΔC , in the supervised targets of GDFM corrected by the CDFM output. In Method B, ΔC is estimated by CDFM in terms of temperature. However the constructed CDFM does not consider that the temperature only dominates approximately 40% of the residential consumption behavior [33], and that the CDFM also needs to eliminate the influence of ΔG . In Method C, continuous time lags sequences of estimated historical consumption calculated by the estimated PV generation of GDM and temperature data are further added as features in the construction of CDFM, to model the continuity of electricity consumption behaviors and the cumulative effect [34]. The time step of the LSTM is set to 7 of CDFM in this experiment. To resolve the latter defect of failure to eliminate the influence of ΔG , GDFM is invoked to calculate ΔG caused by the irradiation difference on the pairwise net load to correct the supervised targets of the CDFM. GDFM and CDFM correct their own supervised targets with mutual outputs until obtaining "purer" model in a cross-iteration fashion. The input features of the two models are summarized in Table III.

TABLE III
INPUT FEATURES AND OUTPUT OF GDFM AND CDFM

Model	Input Features	Output
GDFM (Method A)	DNI, DHI, hour angle of the sun and day of the year	Generation difference
GDFM (Methods B and C)		Generation difference with ΔC elimination
CDFM (Method B)	Ambient temperature, day of the year, day of the week and month of year	Consumption difference
CDFM (Method C)	Ambient temperature, day of the year, day of the week, month of year, and continuous time lags sequences of estimated historical consumption	Consumption difference with ΔG elimination
GDM	DNI, DHI, hour angle of the sun, and day of the year	Disaggregation generation

Fig. 4 gives a comparison of the test error of compensatory consumption difference ΔC of GDFM and compensatory generation difference ΔG of CDFM of the three proposed methods. The number of the iteration for the GDFM model of Method C is 20.



(a) Test error of consumption difference ΔC of GDFM

(b) Test error of generation difference ΔG of CDFM

Fig. 4. Test error of compensatory consumption difference ΔC of GDFM and compensatory generation difference ΔG of CDFM.

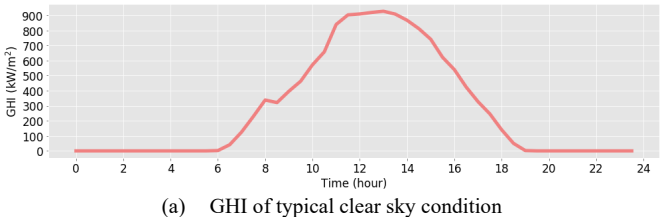
In the left subplot of Fig. 4(a), the test error of consumption difference ΔC is presented. In Method A this appears as a straight line, because there is no correction of ΔC . Comparing to Method B, Method C shows the lowest test error. The first iterations of the red line represent the first iterative update point of Method C, which demonstrates a significant consumption compensation effect, and the error in the subsequent iterations is then maintained in a low range. The reason for the strong explanatory ability of ΔC in Method C of GDFM is that, in addition to the temperature, the CDFM is built using the paired continuous time lags sequences of estimated historical consumption itself. The model mechanism is more like an online forecasting, except that the variable being explained is the load difference rather than the load. In the left subplot of Fig. 4(b), ΔG also decreases significantly through iterative updates and always remains at a low error level. The local iterative results are also presented in the right subplots in Fig. 4(a) and Fig. 4(b), and the gradually decreasing curves trend shows that the subsequent iterative processes are still gradually reducing the test error of GDFM and CDFM, illustrating the effectiveness of cross-iteration mechanism of the proposed method.

In order to analyze the performance of the proposed methods, they are compared with a data-driven based method [24] and a physical model based method [7]. The MSE and CV of various net load disaggregation methods of the real-world dataset are presented in Table IV.

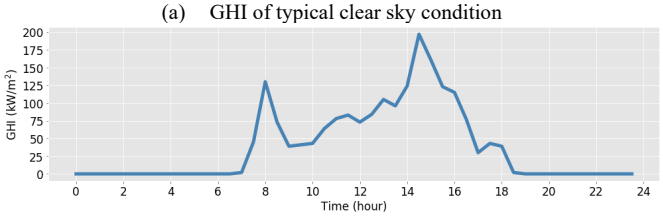
TABLE IV
MSE AND CV OF VARIOUS NET LOAD DISAGGREGATION METHODS

Evaluation metrics	Method [24]	Method [7]	Method A	Method B	Method C
MSE (kW^2)	144.05	101.32	107.01	82.73	**63.77**
CV (%)	9.79	8.70	9.11	8.14	**7.49**

GDM built by Method A has the worst performance due to 1) the assumption of the ideal state of $\Delta C = 0$; 2) lack of correction for the consumption difference. Both GDM model built by Method B and Method C have lower MSE and CV compared to methods in [24] and [7], and Method C has the best performance due to the refinement of the GDFM and CDFM by the cross-iteration on the supervised targets of net load difference. Compared to methods in [24] and [7], Method C exhibits a decreased MSE by 80.28kW^2 and 37.55kW^2, respectively, and a decreased CV by 2.30% and 1.21%, respectively. Corresponding global horizontal irradiation (GHI) and more detailed disaggregation results of the five methods for typical clear sky and non-clear sky conditions are shown in Figs. 5 and 6.

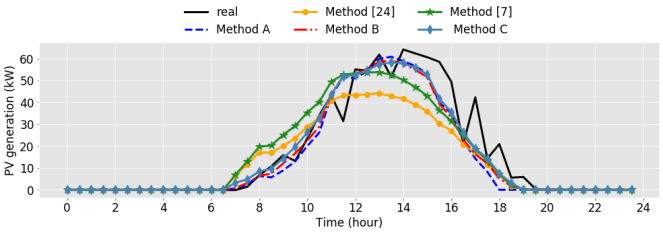


(a) GHI of typical clear sky condition

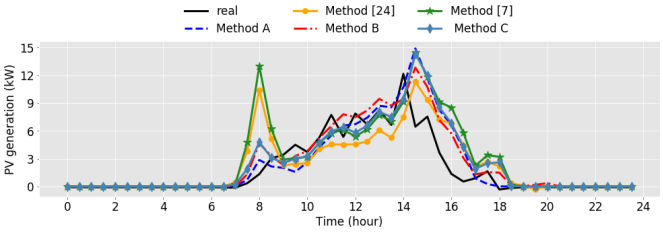


(b) GHI of non-typical clear sky condition

Fig. 5. Irradiation profiles of typical clear sky and non-clear sky conditions.



(a) Disaggregation generation of typical clear sky condition



(b) Disaggregation generation of non-typical clear sky condition

Fig. 6. Disaggregation profiles of typical clear sky and non-clear sky conditions of different methods.

The proposed Method C has the highest disaggregation accuracy even under non-clear sky conditions. It is worth noting that the method in [7] assumes a physical model, which has a fixed shape of the generation curve compared to irradiation curve. The method makes it difficult to reflect the transposition error caused by the distance of the collected meteorological data and the customers to be disaggregated when they are spatially different. The lower disaggregation accuracy of the method in [24] may be caused by the poor explanation ability of its simplified linear formulation to the complex multi-factor driven consumption behavior. Fig. 7 presents a box plot comparison of the disaggregation accuracy of the various proposed methods for each month. It is shown that the proposed Method C has the lowest CV with a particularly significant improvement in disaggregation accuracy in the summer months, when the temperature-sensitive loads are more prevalent.

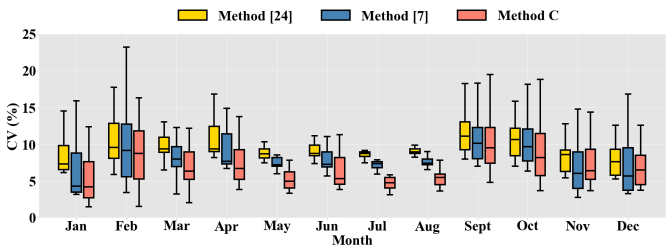


Fig. 7. Disaggregation accuracy of the proposed methods for each month.

Since the specific geographic coordinates of the customers are unknown and the meteorological data used are from a certain weather station in Austin, it is likely that the PV generation data of customers and the solar radiation data collected by the weather station are not exactly matched. This can also explain that the proposed method can have satisfied disaggregation results even using adjacent meteorological data, which can be the case in practical scenarios.

To further investigate the robustness of the proposed method, normal distribution noise is added to the original net load data for each data point. The maximum value of net load data is 96.72kW. The mean of the normal distribution error is set from -10kW to 10kW at intervals of 5kW, and the standard deviation is set from 0% to 50% of the maximum value of the original net load at intervals of 10% (when the standard deviation is set to 0, the normal distribution error can be seen as uniform distribution error). The MSE and CV results of the proposed method with different normal distribution noise additions are shown in Tables V and VI respectively.

TABLE V
MSE (kW^2) WITH DIFFERENT NORMAL DISTRIBUTION NOISE ADDITIONS

Standard deviation (percentage of the maximum net load value)	Mean				
	-10kW	-5kW	0kW	5kW	10kW
0%	62.41	64.09	63.77	63.51	63.74
10%	55.00	57.05	65.61	57.60	53.58
20%	61.26	68.75	65.45	62.23	60.12
30%	82.84	71.81	67.74	62.96	66.25
40%	61.98	68.83	64.81	63.55	97.37
50%	95.87	74.38	63.88	83.19	74.70

TABLE VI
CV (%) WITH DIFFERENT NORMAL DISTRIBUTION NOISE ADDITIONS

Standard deviation (percentage of the maximum net load value)	Mean				
	-10kW	-5kW	0kW	5kW	10kW
0%	7.64	7.64	7.49	7.55	7.58
10%	7.87	7.70	7.80	7.66	7.73
20%	8.30	7.82	7.81	7.60	8.10
30%	8.54	8.89	8.28	8.51	8.45
40%	8.52	8.04	8.83	8.99	8.68
50%	10.80	8.84	7.81	9.55	8.29

From Tables V and VI, it can be seen that in general there is a slight increase in the disaggregation MSE and CV of the

proposed method when the mean value of the added error distribution is fixed and the standard deviation gradually increases, except in the results of the added error of standard deviation 40% mean -5kW, standard deviation 50% mean 0kW and standard deviation 50% mean 10kW. The disaggregation results of Tables V and VI also reflect an intriguing phenomenon that the disaggregation performance of the proposed method has less change at the same standard deviation, regardless of the mean value of the error. This is particularly evident when the standard deviation is zero (uniform distribution error). This is because the method is based on the PV differences and consumption differences obtained from the matched net load differences. The kernel of the modelling essentially reflects the relationship between the changes in the dependent variable difference caused by the changes in the independent variable difference. Thus, the error of the same distribution will be substantially offset in the net load difference processing (this can also be used to explain the occasional phenomenon in Table VI that the disaggregation results become better as the added error increases). This shows the high robustness of the proposed algorithm. This is particularly effective in cases there is an overall net load uniform distribution deviation caused by poor meter pointer calibration.

D. Scenario of Different DPVS Penetration

Different PV penetration causes different degrees of alteration in the net load curve. The net load curves with high penetration DPVS better reflects the morphological features of PV generation, while it is more likely that the net load curves with low penetration hide the energy fluctuations caused by PV generation. To investigate the disaggregation performance of each method at different PV penetration levels, semi-synthetic data (PV penetration of 30%, 60%, 90%, 120% and 150%) are produced by SAM using meteorological data of NSRDB, and the penetration of PV is calculated based on the ratio between total generation and consumption [35]. Disaggregation results at different penetration levels are presented in Tables VII and VIII.

TABLE VII
MSE (kW^2) OF DIFFERENT PV PENETRATION

Penetration	Method [24]	Method [7]	Method A	Method B	Method C
30%	21.15	10.97	18.80	3.27	**1.79**
60%	65.43	44.92	25.45	5.84	**5.13**
90%	137.14	95.12	28.72	14.86	**13.02**
120%	234.90	174.36	36.54	27.27	**18.04**
150%	359.06	271.19	49.08	44.87	**28.26**

TABLE VIII
CV (%) OF DIFFERENT PV PENETRATION

Penetration	Method [24]	Method [7]	Method A	Method B	Method C
30%	10.19	6.80	9.51	5.28	**3.23**
60%	8.88	6.43	6.06	3.67	**3.20**
90%	8.53	6.87	4.70	3.56	**3.07**
120%	8.35	7.14	4.02	3.82	**2.79**
150%	8.25	7.10	3.68	3.76	**2.80**

The proposed Method C exhibits the lowest CV and MSE at all PV penetrations of the experiment, indicating superior performance at different scenarios of PV penetrations. Interestingly, there is not much improvement over Method B as compared to Method A at high PV penetration due to the small values of ΔC , as demonstrated in Fig. 2(a). Therefore, Method A may be a simple and efficient method to disaggregate BTM generation at high penetration levels.

E. Scenario of Different Manufacturing Processes and Technologies

Solar cells can be classified into three generations of manufacturing processes. Currently only the first generation technologies of crystalline silicon wafer-based cells and the second generation technologies of thin-film are used in residential installations [36]. To improve the economic efficiency of DPVS installation, the mainstream practice is to 1) install bifacial modules to improve irradiation absorption or 2) install tracking systems to increase the time of normal irradiation. Both technologies are increasingly popular. Since 2015, about 70% of newly installed utility-scale PV systems have implemented solar tracking [37], [38], while bifacial modules are expected to reach a 40% market share [39].

Differences in manufacturing processes or manufacturers can lead to variations in the solar module characteristics e.g., nominal efficiency, temperature coefficients, etc. and irradiation reception of the PV panels. In order to test the applicability of the proposed algorithm in multiple scenarios, the disaggregation study was conducted on PV systems of mono-c-Si, multi-c-Si, thin film, bifacial mono-c-Si, mono-c-Si with tracking and hybrid. The desired capacity of all the six PV systems is set to 102.53kWp, namely the same nameplate capacity of the real-world data. The generation data of the hybrid system comprises the addition of one-fifth of the generation data of each of the first five PV systems. Daily generation results profiles of the six types of PV systems are shown in Fig. 8.

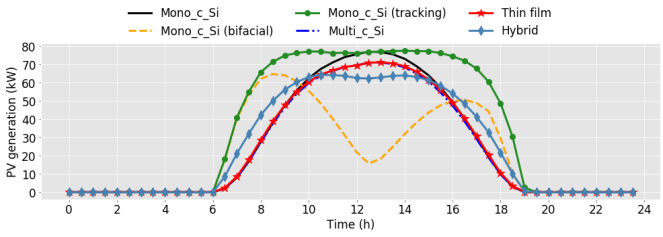


Fig. 8. Daily generation results profiles of six types of PV systems.

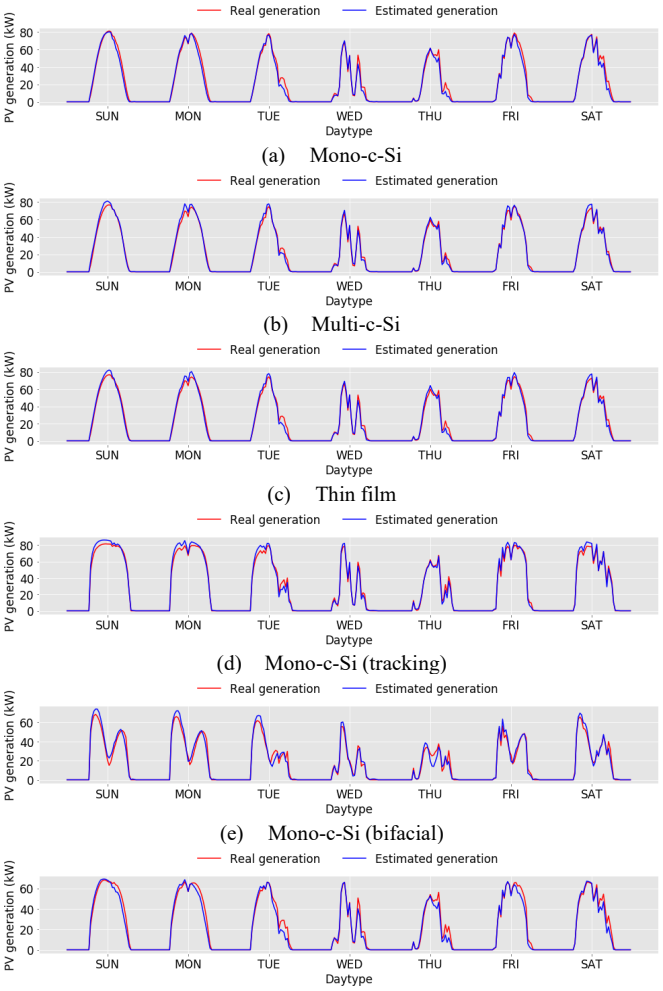
Fig. 8 shows that the PV system of monofacial Mono-c-Si, Multi-c-Si and thin film without tracking exhibit a conventional generation form. In order to fully demonstrate the wide applicability of the proposed method, PV system of bifacial is set to be installed East-West vertical and South-facing tilted, which has been proven to have the optimal generation efficiency in certain scenarios [38]. The generation characteristic of this installation configuration appears as a bimodal state. The PV system with tracking has a longer peak duration because the PV panels are, as much as possible, orthogonal to the irradiation. Since the hybrid PV system is a combination of multiple PV systems, the generation profiles present its own unique form. The disaggregation results of different systems for a week are presented in Fig. 9.

It is shown that the proposed method has a good performance of disaggregation for special scenarios, such as tracking,

bifacial and hybrid systems, and the corresponding flat-peaked, bimodal and mixed patterns are well reflected. The numerical disaggregation results of the six PV systems for a whole year are shown in Table IX.

TABLE IX
NUMERICAL DISAGGREGATION RESULTS OF THE SIX PV SYSTEMS FOR A WHOLE YEAR

Type of PV panels	Evaluation Metrics	
	MSE (kW^2)	CV (%)
Mono-c-Si	21.75	2.93
Multi-c-Si	10.74	2.32
Thin film	15.15	2.51
Mono-c-Si (tracking)	11.41	2.06
Mono-c-Si (bifacial)	21.38	4.04
Hybrid	26.99	3.39



(a) Mono-c-Si
(b) Multi-c-Si
(c) Thin film
(d) Mono-c-Si (tracking)
(e) Mono-c-Si (bifacial)
(f) Hybrid

Fig. 9. Disaggregation results of six types of PV systems for a week.

The disaggregation accuracy for various PV systems proves that the proposed method has a strong general applicability of the BTM PV disaggregation problem, even for hybrid energy systems. Its data-driven nature gives it the ability to disaggregate net load without prior knowledge of the corresponding configurations, e.g., installation number of PV systems, type of PV systems, etc., and the combination of PV systems, which is not possible with model-based disaggregation methods.

IV. CONCLUSIONS

A novel data-driven method for BTM PV disaggregation based on GDFM and CDFM without relying on any energy proxy is proposed in this paper. The proposed algorithm has the following advantages:

1) The proposed method can effectively reduce errors caused by model assumptions that do not conform to actual conditions. Besides, their energy proxy independent characteristic enables the proposed method to effectively avoid the transposition error introduced by the spatial difference and the mismatched energy output characteristics between the target system and proxies.
2) The innovative conversion of GDFM to GDM makes it possible to use supervised learning without knowing the energy data of separate metering of PV and load. By using deep LSTM, the effects of generation difference and consumption difference caused by meteorological data and historical net load data can be captured when building GDFM and CDFM, respectively; the cross-iteration update further refines the disaggregation process.
3) The performance of the proposed BTM generation disaggregation method under various penetration scenarios and PV system types using real-world data is satisfactory, proving its practical application value.

Future work will focus on the disaggregation of BTM energy systems with various technologies, such as energy storage and electric vehicles. The extensibility of the proposed model allows the possibility of exploring higher disaggregation accuracy utilizing more advanced supervised learning methods. We also have an interest to infer the PV installation capacity while estimating the PV generation disaggregation results, and in exploring the performance of the application in case of high resolution metering data.

APPENDIX
HYPERPARAMETERS OF GDFM AND CDFM

Both the GDFM and CDFM are optimized by Adam algorithm [40]. The hyperparameters of GDFM and CDFM are listed in Table X.

TABLE X
HYPERPARAMETERS OF GDFM AND CDFM

Model	Hidden layers	Time step of LSTM	Batch size	Epochs[b]
GDFM (Methods A, B and C)	150[a]-150[a]-150-100-50	1	4096	50-20
CDFM (Method B)	100-20	/	4096	50
CDFM (Method C)	Input branch 1: 150[a]-150[a]-150[a] Input branch 2: 50 Concat[c]: 100-20	7	4096	30-10

[a]: The units are LSTM units and the others are dense units
[b]: The first number represents the training epoch at the first iteration; the second number represents the training epoch when the number of iterations is greater than 1, and the weight of new iteration is initialized by the weight saved at the end of the previous iteration.
[c]: Concat means the concatenates of outputs of Input branch 1 and Input branch 2.

REFERENCES

[1] H. L. van Soest, M. G. J. den Elzen, and D. P. van Vuuren, "Net-zero emission targets for major emitting countries consistent with the Paris Agreement," Nat. Commun., vol. 12, no. 1, pp. 1–9, Apr, 2021.

[2] C. S. Lai et al., "Levelized cost of electricity for photovoltaic/biogas power plant hybrid system with electrical energy storage degradation costs," Energy Convers. Manag., vol. 153, pp. 34–47, Dec, 2017.

[3] "Solar PV." https://www.iea.org/reports/renewables-2020/solar-pv (accessed Jul. 17, 2021).

[4] K. S. Hayibo and J. M. Pearce, "A review of the value of solar methodology with a case study of the U.S. VOS," Renew. Sustain. Energy Rev., vol. 137, pp. 110599, Apr, 2021.

[5] J. Brown, A. Abate, and A. Rogers, "Disaggregation of household solar energy generation using censored smart meter data," Energy Build., vol. 231, pp. 110617, Jan, 2021.

[6] E. Liu, S. Member, and P. Cheng, "Achieving privacy protection using distributed load scheduling : A randomized approach," IEEE Trans. Smart Grid, vol. 8, no. 5, pp. 2460–2473, May, 2017.

[7] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV," IEEE Trans. Power Syst., vol. 33, no. 3, pp. 3255–3264, 2018.

[8] F. Wang, K. Li, C. Liu, Z. Mi, M. Shafie-Khah, and J. P. S. Catalao, "Synchronous pattern matching principle-based residential demand response baseline estimation: Mechanism analysis and approach description," IEEE Trans. Smart Grid, vol. 9, no. 6, pp. 6972–6985, Oct, 2018.

[9] X. Zhang and S. Grijalva, "A data-driven approach for detection and estimation of residential PV installations," IEEE Trans. Smart Grid, vol. 7, no. 5, pp. 2477–2485, Apr, 2016.

[10] F. Shen, Q. Wu, J. Zhao, W. Wei, N. D. Hatziargyriou, and F. Liu, "Distributed risk-limiting load restoration in unbalanced distribution systems with networked microgrids," IEEE Trans. Smart Grid, vol. 11, no. 6, pp. 4574–4586, May, 2020.

[11] F. Ding and B. Mather, "On distributed PV hosting capacity estimation, sensitivity study, and improvement," IEEE Trans. Sustain. Energy, vol. 8, no. 3, pp. 1010–1020, Dec, 2017.

[12] H. Shaker, H. Zareipour, and D. Wood, "Estimating power generation of invisible solar sites using publicly available data," IEEE Trans. Smart Grid, vol. 7, no. 5, pp. 2456–2465, Mar, 2016.

[13] H. Shaker, D. Manfre, and H. Zareipour, "Forecasting the aggregated output of a large fleet of small behind-the-meter solar photovoltaic sites," Renew. Energy, vol. 147, pp. 1861–1869, Mar, 2020.

[14] E. C. Kara, C. M. Roberts, M. Tabone, L. Alvarez, D. S. Callaway, and E. M. Stewart, "Disaggregating solar generation from feeder-level measurements," Sustain. Energy, Grids Networks, vol. 13, pp. 112–121, Mar, 2018.

[15] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Duić, and T. Wang, "Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation," Appl. Energy, vol. 253, pp. 113595, Nov, 2019.

[16] C. M. Cheung, W. Zhong, C. Xiong, A. Srivastava, R. Kannan, and V. K. Prasanna, "Behind-the-meter solar generation disaggregation using consumer mixture models," 2018 IEEE Int. Conf. Commun. Control. Comput. Technol. Smart Grids, SmartGridComm 2018, 2018.

[17] C. M. Cheung, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna, "Disaggregation of behind-the-meter solar generation in presence of energy storage resources," 2020 IEEE Conf. Technol. Sustain. SusTech 2020, no. 1911229, 2020.

[18] M. Tabone, S. Kiliccote, and E. C. Kara, "Disaggregating solar generation behind individual meters in real time," BuildSys 2018 - Proc. 5th Conf. Syst. Built Environ., pp. 43–52, 2018.

[19] M. Khodayar, G. Liu, J. Wang, O. Kaynak, and M. E. Khodayar, "Spatiotemporal behind-the-meter load and PV power forecasting via deep graph dictionary learning," IEEE Trans. Neural Networks Learn. Syst., pp. 1–15, Dec, 2020.

[20] H. Shaker, H. Zareipour, and D. Wood, "A data-driven approach for estimating the power generation of invisible solar sites," IEEE Trans. Smart Grid, vol. 7, no. 5, pp. 2466–2476, Dec, 2015.

[21] F. Bu, K. Dehghanpour, Y. Yuan, Z. Wang and Y. Zhang, "A data-driven game-theoretic approach for behind-the-meter PV generation disaggregation," IEEE Trans. Power Syst., vol. 35, no. 4, pp. 3133–3144, Jan, 2020.

[22] W. Li, M. Yi, M. Wang, Y. Wang, D. Shi, and Z. Wang, "Real-time energy disaggregation at substations with behind-the-meter solar generation," IEEE Trans. Power Syst., vol. 36, no. 3, pp. 2023–2034, Nov, 2020.

[23] W. Stainsby, D. Zimmerle, and G. P. Duggan, "A method to estimate residential PV generation from net-metered load data and system install date," Appl. Energy, vol. 267, pp. 114895, Nov, 2020.

[24] F. Sossan, L. Nespoli, V. Medici, and M. Paolone, "Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers," IEEE Trans. Ind. Informatics, vol. 14, no. 9, pp. 3904–3913, Sept, 2018.

[25] D. Chen and D. Irwin, "SunDance: Black-box behind-the-meter solar disaggregation," e-Energy 2017 - Proc. 8th Int. Conf. Futur. Energy Syst., pp. 45–55, 2017.

[26] F. Kabir, N. Yu, W. Yao, R. Yang, and Y. Zhang, "Estimation of behind-the-meter solar generation by integrating physical with statistical models," 2019 IEEE Int. Conf. Commun. Control. Comput. Technol. Smart Grids, SmartGridComm 2019, 2019.

[27] F. Kabir, N. Yu, W. Yao, R. Yang, and Y. Zhang, "Joint estimation of behind-the-meter solar generation in a community," IEEE Trans. Sustain. Energy, vol. 12, no. 1, pp. 682–694, Jan, 2021.

[28] K. Pan, C. Xie, C. S. Lai, D. Wang, and L. L. Lai, "Photovoltaic output power estimation and baseline prediction approach for a residential distribution network with behind-the-meter systems," Forecasting, vol. 2, no. 4, pp. 470–487, Nov, 2020.

[29] "PECAN STREET|DATAPORT." https://dataport.pecanstreet.org/# (accessed Jul. 17, 2021).

[30] "NSRDB Data Viewer." https://maps.nrel.gov/nsrdb-viewer/?aL=x8CI3i%255Bv%255D%3Dt%26ozt_aP%255Bv%255D%3Dt%26ozt_aP%255Bd%255D%3D1&bL=clight&cE=0&lR=0&mC=4.740675384778373%2C22.8515625&zL=2 (accessed Jul. 17, 2021).

[31] "System Advisor Model (SAM)." https://sam.nrel.gov/download.html (accessed Jul. 17, 2021).

[32] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 841–851, Jan, 2019.

[33] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," Energy Build., vol. 40, no. 3, pp. 394–398, 2008.

[34] T. Teeraratkul, D. O'Neill, and S. Lall, "Shape-based approach to household electric load curve clustering and prediction," IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 5196–5206, Sept, 2018.

[35] B. Johnson, P. Denholm, B. Kroposki, and B. Hodge, "Achieving a 100% renewable grid," IEEE Power Energy Mag., no 2, pp. 61–73, Apr, 2017.

[36] M. H. Shubbak, "Advances in solar photovoltaics: Technology review and patent trends," Renew. Sustain. Energy Rev., vol. 115, pp. 109383, Jul, 2019.

[37] A. Onno et al., "Predicted power output of silicon-based bifacial tandem photovoltaic systems," Joule, vol. 4, no. 3, pp. 580–596, Mar, 2020.

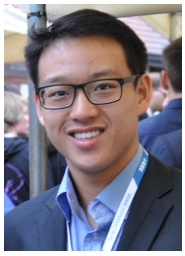
[38] M. T. Patel, M. S. Ahmed, H. Imran, N. Z. Butt, M. R. Khan, and M. A. Alam, "Global analysis of next-generation utility-scale PV: Tracking bifacial solar farms," Appl. Energy, vol. 290, pp. 116478, Mar, 2021.

[39] VDMA, "International Technology Roadmap for Photovoltaic," Itrpv, vol. 11th Editi, pp. 76, Apr, 2020, [Online]. Available: https://itrpv.vdma.org/en/ueber-uns.

[40] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

Keda Pan received the B.Eng. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2015 and the M.Eng. degree in electrical engineering from the Guangdong University of Technology, Guangdong, China, in 2019, respectively. He is currently working toward the Ph.D. degree in control science and engineering with the Guangdong University of Technology, Guangdong, China. His research interests include data analytics and demand response programs in behind-the-meter energy system.

Zhaohua Chen received the B.Eng. degree in electrical engineering and automation from the Dongguan University of Technology, Dongguan, China, in 2019. He is currently working toward the master's degree in electrical engineering with the School of Automation, Guangdong University of Technology, Guangzhou, China. His research interests include deep learning and its applications in power systems.



Chun Sing Lai (Senior Member, IEEE) received the B.Eng. (First Class Hons.) in electrical and electronic engineering from Brunel University London, London, UK, in 2013, and the D.Phil. degree in engineering science from the University of Oxford, Oxford, UK, in 2019.

He is currently a Lecturer with the Department of Electronic and Electrical Engineering, Brunel University London. From 2018 to 2020, he was an UK Engineering and Physical Sciences Research Council Research Fellow with the School of Civil Engineering, University of Leeds, Leeds, UK. Dr Lai was the Publications Co-Chair for both 2020 and 2021 IEEE International Smart Cities Conferences. He is the Vice-Chair of the IEEE Smart Cities Publications Committee and Associate Editor for IET Energy Conversion and Economics. He is the Working Group Chair for IEEE P2814 Standard, Associate Vice President, Systems Science and Engineering of the IEEE Systems, Man, and Cybernetics Society (IEEE/SMCS), the Chair of the IEEE SMC Intelligent Power and Energy Systems Technical Committee and Technical Program Chair of the 2022 IEEE Smart Cities Conference. As the lead author, he has contributed to three journal articles that appeared in Web of Science as Highly Cited Papers. He is an IET Member and a Chartered Engineer. His current research interests are in power system optimization and data analytics.

Changhong Xie received the B.Eng. degree in electrical engineering and automation from Nanchang Institute of Technology, Nanchang, China, in 2019. He is currently pursuing the M.Eng. degree in electrical engineering with the Guangdong University of Technology. His current research interests include energy storage systems, renewable energy, and data analysis.

Dongxiao Wang (Member, IEEE) received the B.Eng. degree in thermal energy and power engineering from North China Electric Power University, China, in 2014, and the D.Phil. degree in electrical engineering from the University of Newcastle, Australia, in 2017. He is currently a Senior Analyst at Australia Energy Market Operator, Australia. His current research interests include demand response technologies, smart energy management, and big data analysis.

Zhuoli Zhao (Member, IEEE) received the Ph.D. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2017. From October 2014 to December 2015, he was a Joint Ph.D. Student and Sponsored Researcher with the Control and Power Research Group, Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. He was a Research Associate with the Smart Grid Research Laboratory, Electric Power Research Institute, China Southern Power Grid, Guangzhou, China, from 2017 to 2018. He is currently an Associate Professor with the School of Automation, Guangdong University of Technology, Guangzhou, China. His research interests include microgrid control and energy management, renewable power generation control and grid-connected operation, modeling, analysis and control of power-electronized power systems and smart grids. He is an Active Reviewer for the IEEE Transactions on Smart Grid, the IEEE Transactions on Sustainable Energy, the IEEE Transactions on Industrial Electronics, the IEEE Transactions on Power Electronics, and the Applied Energy.

Xiaomei Wu (Member, IEEE) received the B.Eng. and M.Eng. degrees in electrical engineering from Northeast Institute of Electric Power, Jilin, China, in 1993 and 1997, respectively. She also received the M.Eng. degree in electrical engineering from Concordia University, Montreal, Canada, in 2004 and the Ph.D. degree in control science and engineering from Guangdong University of Technology, Guangzhou, China, in 2021. From 2009 up to now, she was an associate professor with the Guangdong University of Technology, Guangzhou, China. Her research interest are hybrid renewable energy systems, electric vehicles, power system operation, and power markets.

Ning Tong received a Ph.D. degree in electrical engineering from Huazhong University of Science and Technology in 2016. He is an assistant professor at the School of Automation, Guangdong University of Technology, China. His research interests include novel protection relays, VSC-HVDC, and renewable energy integration.

Loi Lei Lai (Life Fellow, IEEE) received the B.Sc. (First Class Hons.), Ph.D., and D.Sc. degrees in electrical and electronic engineering from the University of Aston, Birmingham, UK, and City, University of London, London, UK, in 1980, 1984, and 2005, respectively. Professor Lai is currently a University Distinguished Professor with Guangdong University of Technology, Guangzhou, China. He was a Pao Yue Kong Chair Professor with Zhejiang University, Hangzhou, China, and the Professor and Chair of Electrical Engineering with City, University of London. His current research areas are in smart cities and smart grid. Professor Lai was awarded an IEEE Third Millennium Medal, the IEEE Power and Energy Society (IEEE/PES) UKRI Power Chapter Outstanding Engineer Award in 2000, a special award from City, University of London in 2005 and is its honorary graduate, the IEEE/PES Energy Development and Power Generation Committee Prize Paper in 2006 and 2009, the IEEE/SMCS Outstanding Contribution Award in 2013 and 2014, the Most Active Technical Committee Award in 2016, and his research team received a Best Paper Award in the IEEE International Smart Cities Conference in October 2020. Professor Lai is an Associate Editor of the IEEE Transactions on Systems, Man, and Cybernetics: Systems, Editor-in-Chief of the IEEE Smart Cities Newsletter, a member of the IEEE Smart Cities Steering Committee and the Chair of the IEEE Systems, Man, and Cybernetics Society (IEEE/SMCS) Standards Committee. He was a member of the IEEE Smart Grid Steering Committee; the Director of Research and Development Center, State Grid Energy Research Institute, China; a Vice President for Membership and Student Activities with IEEE/SMCS; a Fellow Committee Evaluator for the IEEE Industrial Electronics Society and IEEE PES Lifetime Achievement Award Assessment Committee Member. He is a Fellow of the IET.

Nikos D. Hatziargyriou (Life Fellow, IEEE) is a Professor Emeritus of Power Systems with the National Technical University of Athens. He has over ten year industrial experience as the Chairman and the CEO of the Hellenic Distribution Network Operator and as the Executive Vice-Chair and the Deputy CEO of the Public Power Corporation, responsible for the Transmission and Distribution Divisions. He was the Chair and the Vice-Chair of the EU Technology and Innovation Platform on Smart Networks for Energy Transition. He is an Honorary Member of CIGRE and the Past Chair of CIGRE SC C6 "Distribution Systems and Distributed Generation." He has participated in more than 60 research and development projects funded by the EU Commission, electric utilities and industry for fundamental research and practical applications. He is the author of the book Microgrids: Architectures and Control and of more than 250 journal publications and 500 conference proceedings papers. He is the 2017 recipient of the IEEE/PES Prabha S. Kundur Power System Dynamics and Control Award. He is the Past Chair of the Power System Dynamic Performance Committee and the Editor-in-Chief of the IEEE Transactions on Power Systems. He is included in the 2016, 2017, and 2019 Thomson Reuters lists of the top 1% most cited researchers and he is the 2020 Globe Energy Prize Laureate.