Clustering-Based Residential Baseline Estimation: A Probabilistic Perspective

Mingyang Sun, *Member, IEEE,* Yi Wang, *Student Member, IEEE,* Fei Teng, *Member, IEEE,* Yujian Ye, *Member, IEEE,* Goran Strbac, *Member, IEEE,* Chongqing Kang, *Fellow, IEEE*

***Abstract*—Demand Response (DR) is one of the most cost-effective solutions for providing flexibility to power systems. The extensive deployment of DR trials and the roll-out of smart meters enable the quantification of consumer responsiveness to price signals via baseline estimation. The traditional deterministic baseline estimation approach can provide only a single value without consideration of uncertainty. This paper proposes a novel probabilistic baseline estimation framework that consists of a daily load profile pool construction stage, a deep learning-based clustering stage, an optimal cluster selection stage, and a quantile regression forests model construction stage. In particular, the concept of a daily load profile pool is introduced, and a deep-learning-based clustering approach is employed to handle a large number of daily patterns to further improve the baseline estimation performance. Case studies have been conducted on fine-grained smart meter data collected from a real dynamic time-of-use (dTOU) tariffs trial of the Low Carbon London (LCL) project. The superior performance of the proposed method is demonstrated based on a series of evaluation metrics regarding both deterministic and probabilistic estimation results.**

***Index Terms*—Deep learning, demand response, probabilistic baseline estimation, clustering, dynamic time-of-use tariff.**

NOMENCLATURE

Sets and indices

Ω	Set of all days, indexed j
Ω^B	Set of days without dTOU events, indexed j
Ω^E	Set of days with dTOU events, indexed j

Variables

$k_{i,j}^{opt}$	Index of the optimal cluster for day j of customer i
X_i^{Train}/X_i^{Test}	Training and test features of model F_i
Y_i^{Train}/Y_i^{Test}	Training and test labels of model F_i (kW)

Functions

DP	The dropout function
F_i	Quantile regression forests model for customer i

Input Parameters

Π	The dTOU tariff data (pence/kWh)
$d_{i,j}^{base}$	The pre- and post-event load data (kW)
D^{dTOU}	Demand measurements of the dTOU group (kW)
$d_{i,j}^{event}$	The DR event data (kW)
D^{nonTOU}	Demand measurements of the nonTOU group (kW)
$d_{i,j}$	The load data of day j, customer i (kW)
K	Number of clusters
M_d	Number of the dTOU customers
M_n	Number of the nonTOU customers
N_k	Number of daily load patterns in cluster k
N_n	Total number of daily load patterns
N_V	Number of neurons for each layer
T	Total number of data points
t	Number of data points within a day
V	Number of layers

M. Sun, F. Teng, Y. Ye and G. Strbac are with the Department of Electrical & Electronic Engineering, Imperial College London, London, SW7 2AZ, UK.

Y. Wang and Chongqing Kang are with the State Key Lab of Power Systems, Dept. of Electrical Engineering, Tsinghua University, Beijing 100084, China.

I. INTRODUCTION

DEMAND response (DR) is one of the most flexible and effective solutions to reduce system operation costs and displace generation and network reinforcement [1]. In general, DR can be classified as either price-based or incentive-based [2]. Specifically, incentive-based programs, such as direct load control and interruptible service, offer fixed or time-varying incentives to customers for their corresponding demand changes. Price-based DR directly sends variational price signals to customers in order to incentivize modification of the customer's electricity consumption. For the price-based DR, there are different types of dynamic pricing schemes, such as day-ahead pricing, real-time pricing, critical peak pricing and dTOU tariffs [3].

One of the key barriers for DR programme is to quantify the responsiveness of users, which is critical for flexibility aggregation, tariff settlement and design [4], [5]. Although the measured electricity consumption during dTOU events are often available, it is challenging to obtain the accurate estimation of the baseline electricity consumption assuming DR events have not happened. For incentive-based DR program, the customer baseline load (CBL) is the basis to determine the compensations given to DR participants for their load reduction during DR event. Inaccurate baseline estimation results in unfairness between the DR participants and aggregators. As illustrated in [6], overestimated baseline demand may cause aggregators to offer more compensation to customers, whereas underestimation will influence customer responsiveness due to the insufficient rewards. On the other hand, for price-based DR program (e.g., dTOU), the CBL estimation is performed to evaluate the DR performance of the dTOU program and quantify the customer responsiveness, which is the main motivation of this work. The conventional baseline estimation approaches presented in the literature can be classified into *similar-day-based methods*, *regression-based methods* and *morning-consumption-adjustments-based methods* [7]. Similar-day-based methods [8] (e.g., a simple ten-day average or the average over several selected days) are typically

carried out by selecting and averaging a series of similar days in terms of the weather conditions to estimate the baseline during the event periods. Regression-based methods (e.g., exponential smoothing model [9]) are typically performed based on historical data to "forecast" the baseline demand during event periods. However, this type of approach yields inaccurate estimation because very limited observations are available for extreme weather days, in which DR events are usually designed to occur. To improve the estimation accuracy, morning usage adjustment methods, which employ pre-event data to adjust the estimated baseline demand during event periods, were proposed in [10].

In recent years, the influx of fine-grained data provides a valuable opportunity to improve the baseline estimation performance. For example, the authors in [11] propose a novel control group selection method to estimate the residential baseline for different DR periods or several events within the same day. Compared with the traditional similar-day-matching and regression-based methods, the control group selection method can protect the estimation from the influence of the anticipation effect and the requirement of a large amount of historical data. Furthermore, a series of clustering-based methods have been proposed in the literature to group the residential load profiles and identify representative load profiles for baseline estimation. In [12], a two-stage adaptive baseline estimation method that leverages the self-organizing map (SOM) and k-means clustering methods is proposed to identify the days most similar to the tested day under DR events. The k-means clustering method is also implemented in [7] to group the average daily load profiles based on real datasets from a utility company. In [6], the authors propose a novel synchronous-pattern-matching-principle-based baseline estimation approach to address the non-synchronous matching issues in the previous study. Although improvements have been demonstrated in terms of their estimation accuracy, it is imperative to note that most of the current research focuses on deterministic baseline estimation methods, which are unable to characterize uncertainties inherent in a distribution network and arising from various consumer behaviors. As illustrated in [13], deterministic methods fail to utilize the historical data in capturing the dynamics of complex user behaviors, particularly important for small to medium consumers with more variability. In other words, due to the massive uncertainty associated with the response from residential customers, the motivation to propose a probabilistic baseline estimation in this work is to provide the customer responsiveness with a confidence interval of DR at each time step per customer. This can help the aggregators with different risk preferences to better understand their customers' capability and certainty of DR and then can be used for residential customer load control, DR targeting, tariff design and/or bidding under stochastic/robust optimization with different levels of risk.

To this end, the authors in [13] and [14] propose a probabilistic baseline estimation approach by employing Gaussian process regression. As was shown, the proposed data-driven method can successfully identify customer load patterns with probabilistic estimation. Recently, an empirical probabilistic baseline estimation approach based on quantile regression (QR) was introduced in [15] to further characterize residential electricity consumer responsiveness to incentives using real data from the ADDRESS project. Note that the k-means clustering method is also employed in this work but based on consumer flexibility. Specifically, clustering is performed based on the estimated parameters of the individual QR models. However, the performance of the QR-based approach can be significantly restricted because of the limitations of k-means (e.g., randomized initial centroids leading to different clustering results, limited performance when dealing with clusters of different density).

In this paper, we propose a novel probabilistic baseline estimation method that uses a deep learning-based clustering method. A daily load profile pool of non-dTOU customers is constructed without consideration of different load conditions and provides sufficient daily load patterns for deep learning to extract effective representative features for clustering. Afterwards, optimal cluster selection is performed for each tested event day of each customer, and then quantile regression forests models are constructed to generate the probabilistic estimated baseline demand during dTOU events. Case studies are conducted based on the real smart meter data from the LCL dTOU trial. The superior performance of the proposed framework is demonstrated by comparing it with other existing methods in terms of the considered evaluation metrics. To summarize, the contributions of this paper are as follows:

1) A novel probabilistic baseline estimation framework is designed to quantify electricity consumer responsiveness at the household level while capturing the uncertainty associated with residential customers. In particular, the proposed framework fully utilizes the synchronous information from nonTOU customers and leverage both pre-event and post-event data of the tested dTOU customer. Instead of building and fixing the optimal clusters over the whole period, the proposed method aims to select the most similar patterns for each day of each individual tested dTOU customer. The concept of a daily load profile pool is proposed to significantly enrich the database of daily load patterns for clustering-based baseline estimation methods. In addition, the challenges of different event durations and start-times are handled in this work.

2) A deep learning-based clustering method, namely, deep embedded clustering (DEC), is employed for the first time to cluster this large number of daily load patterns, integrating dimensionality reduction and clustering into a single end-to-end learning framework. DEC has the advantage of jointly extracting informative features and construct separable clusters.

3) Compared with a series of conventional and state-of-the-art methods, the superior performance of the proposed framework is demonstrated in terms of the point and probabilistic estimation results.

This paper is structured as follows. Section II illustrates the baseline estimation problem. Section III illustrates the proposed baseline estimation framework. Section IV presents the employed technical details. Section V conducts numerical experiments on the LCL dTOU trial dataset. Section VI concludes the paper.

II. Baseline Estimation

As introduced in [16], the definitions of dynamic time-of-use (dTOU) events, non-Time-Of-Use (nonTOU) events, dTOU group, and nonTOU group are given as follows:

1) **dTOU events**: the periods with low price or high price events;
2) **nonTOU events**: the periods with default price;
3) **dTOU group**: the group whose customers received the experimental dTOU tariff;
4) **nonTOU group**: the group whose customers received a standard flat tariff.

DR can be quantified by estimating the change in demand during a dTOU event relative to the baseline demand that would have occurred in the context of a non-dTOU event [16]:

$$DR_e = D_e^{actual} - D_e^{baseline} \tag{1}$$

where D_e^{actual} and $D_e^{baseline}$ represent the actual measurements and the estimated baseline demand during the dTOU event e, respectively. To obtain an accurate estimation of the DR and further investigate the impacts of dTOU tariffs on customer electricity consumption behavior, it is crucial for distribution system operators and utility companies to effectively estimate the baseline demand $D_e^{baseline}$ at the single household level. The conventional deterministic baseline estimation approach may result in erroneous rewards for residential consumers due to their large uncertainties [13]. To this end, given the half-hourly demand measurements of the nonTOU group D^{nonTOU} and the dTOU group D^{dTOU}, as well as the corresponding dTOU tariffs data Π, the target is to estimate probabilistic baseline demand $\hat{D}^{baseline}$ for the dTOU customers via model F, expressed as follows:

$$\hat{D}^{baseline} = F(D^{dTOU}, D^{nonTOU}, \Pi) \tag{2}$$

In terms of the key challenges that pertain to residential baseline estimation, three main aspects can be summarized as follow:

1) ***New task***: Traditional load forecasting refers to predict the expected electricity demand at aggregated levels [17]. With the deployment of smart meter, a massive amount of highly granular data render it possible to forecast the load data at lower levels such as household level [18], which is more challenging due to the high variability and diversity. Under this reality, *the target of probabilistic residential load forecasting is to accurately predict the future electricity demand while capturing the uncertainty arising from diversified consumer behaviors*. Then the generated forecasts can be employed to make decisions for different applications such as system planning, day-ahead scheduling, etc. However, *baseline estimation aims to estimate what the customer would have consumed had the price signal not been sent and then calculating the difference between the estimated baseline and the actual measured load to quantify residential consumer responsiveness* [16]. In addition, probabilistic baseline estimation also needs to capture the uncertainty by outputting density functions, quantiles, or intervals. Therefore, instead of predicting future demand (i.e., load forecasting), baseline estimation is usually conducted after the trail (i.e., post-experiment analysis) based on a sufficient number of actual load measurements for both event and non-event periods. Then the potential value of residential demand can be further investigated when it contributes to system balancing or network reinforcement.

2) ***Time-varying Pricing***: Based on time horizons, load forecasting can be generally grouped into very short term load forecasting (VSTLF) (i.e., $T < 1day$), short term load forecasting (STLF) (i.e., $1day < T < 2weeks$), medium term load forecasting (MTLF) (i.e., $2weeks < T < 3years$), and long term load forecasting (LTLF) (i.e., $3years < T$) [17]. Baseline estimation is usually performed within one day, which corresponds to VSTLF or STLF. Nevertheless, it is imperative to highlight that, unlike load forecasting which has a pre-determined prediction period (e.g., 24 hours), baseline estimation need to deal with the challenge of *time-varying pricing* when dTOU tariffs are considered. For example, in the LCL project, different high price or low price event durations (3 hours, 6 hours, 12 hours, and 24 hours) are designed with various start-times such as 2:00am, 7:00am, 14:00pm, 17:00pm [16]. Consequently, it may be ineffective to employ conventional load forecasting model and features to estimate the baseline demand under such diversified event periods and therefore, it becomes imperative to design new framework and training features for these specific characteristics.

3) ***New features***: Regarding the input features, both load forecasting and baseline estimation consider weather conditions (e.g., temperature, wind speed and humidity [6]) and the time-series information (e.g., hour of the day, day of the week and month of the year) as explanatory variables. However, to estimate the baseline demand for a DR participant, solely based on its historical data renders it difficult to effectively learn the baseline consumption patterns during the dTOU events because they most likely correspond to the extreme weather conditions. In other words, the large amount of fine-grained smart meter data can be sufficient to train an accurate load forecasting model but may be still not enough for baseline estimation. To this end, the CONTROL group methods were proposed in the literature (e.g., [7], [11]) to consider the load data of the non-dTOU customers, who have similar electricity consumption behavior to the dTOU customers, as additional features beyond the weather data and time-series information. Nevertheless, there are still a lot of open questions need to be solved such as *1) What are the appropriate metrics to evaluate the similarity between the non-dTOU customer and the dTOU customer? 2) Because of the high variability of load profiles at household level, it is still not efficient to use the whole data set of the identified similar non-dTOU customer as the training features. Is it possible to design ad-hoc features to further improve the estimation performance?* Additionally, it is important to note that load forecasting can only use the historical data before the test periods whereas baseline estimation can leverage the knowledge learnt from both pre-event and post-event hours to simultaneously consider forward and backward dependencies in time series data.

III. Proposed Framework

The main challenges of residential baseline estimation that will be solved in this paper can be summarized as:

(C1) Fully utilizing the synchronous information from non-TOU customers and leveraging both pre-event and post-event baseline load data of the tested dTOU customer;

(C2) Dealing with time-varying pricing, which leads to different durations and start-times of dTOU events;

(C3) Capturing massive uncertainties of residential baseline demand.

In this work, to deal with the aforementioned challenges, the proposed probabilistic baseline estimation framework shown in Fig.1 consists of the following main stages. In particular, for **(C1)**, the *Daily Load Profile Pool Construction Stage* and the *Deep Learning-Based Clustering Stage* effectively extract the information form nonTOU customers by identifying representative daily baseline load patterns of a pool, consisting of a massive number of possible daily baseline shapes. Then based on the pre-event and post-event baseline load of the tested dTOU customer, the *Optimal Cluster Selection Stage* aims to construct the training and testing features for the *Quantile Regression Forests Model Construction Stage* to build a quantile regression forests model for each customer, which can capture the massive uncertainties of residential baseline demand **(C3)**. With the features of hour of the day, different event durations and start-times **(C2)** can be handled properly. Note that users can make the trade-off between the estimation performance and the model complexity by replacing the DEC and the quantile regression forest with other conventional approaches, which can be easily implemented.

A. Daily Load Profile Pool Construction Stage

As illustrated in [7], clustering-based baseline estimation approaches exhibit the benefits of: (i) inherently incorporating weather, social factors or other exogenous factors into changes in demand by using concurrent electricity consumption data from the control group; (ii) being not limited to similar weather days to render the method feasible for new customers or the events under new abnormal weather conditions.

Nevertheless, conventional clustering-based approaches conduct the group construction based on average load profiles of customers and then fix the established groups over all the event periods. This may result in the following disadvantages. First, solely based on averaging daily load profiles of customers over the test period (e.g., one year), it is difficult to build clusters that can concurrently well distinguish consumer behaviors between clusters and match behavior across all customers within a cluster. Moreover, clustering based on average profiles can not effectively deal with the "outlier" customers (i.e., customers with extremely unusual load patterns). Most importantly, under different weather conditions, clustering results could be different for the same group of customers.

To overcome the aforementioned problems, instead of building and fixing the clusters over the whole trial period, the proposed method aims to select the most similar load patterns for each day of each individual dTOU customer from the representative daily baseline load patterns obtained by conducting clustering on a massive number of daily load profiles of the non-dTOU customers. To this end, the first step is to enrich the daily load profile patterns via a very straightforward but effective way - *Daily Load Profile Pool Construction*, which also improve the utilization of nonTOU customers by extracting more distinctive daily baseline shapes, referring to the challenge **(C1)**.

Let $D^{nonTOU} \in \mathbb{R}^{T\times M_n}$ denote the smart meter data of the nonTOU group customers (i.e., the CONTROL group), where M_n is the number of customers in the nonTOU group and T is the total number of measurements collected during the whole trial period. As presented in Fig. 2, the load profile of the whole trial period T is partitioned into T/t daily load patterns for each nonTOU customer, which are then stacked together one by one to build the output pattern library. The total number of daily load profiles included in the constructed training pool is

$$N_n = \frac{T}{t} \times M_n \tag{3}$$

where t is the number of measurements collected for each day (i.e., $t = 48$ for half-hourly data). It is important to note that in the constructed pool, each daily load profile is normalized by its maximum value. The target of this normalization step is to extract representative daily baseline patterns through the clustering stage without the consideration of the magnitudes. Although the magnitudes of these daily load profiles are not retained in the stage for the nonTOU customers, DR is quantified by calculating the difference between the actual demand and estimated baseline demand for the tested dTOU customers. Therefore, as input features for the regression model, the normalized daily patterns will be scaled based on the magnitudes of the tested dTOU customer during the *Optimal Cluster Selection Stage*. Also, the training label of the regression model is kept in the kW domain so that we can obtain the actual value of load to compute the compensation for these customers as in equation (1).

B. Deep Learning-Based Clustering Stage

The *Deep Learning-Based Clustering Stage* is proposed to cluster the vast number of daily load profiles $D^{train} \in \mathbb{R}^{N_n\times t}$ by employing deep embedded clustering (DEC) technique with the benefits of concurrently extracting informative features and minimizing the clustering loss. Details regarding the DEC approach will be illustrated in Section IV.

Given the number of clusters K, all the N_n daily load profiles can be grouped into K clusters $\{D_k\}_{k=1}^{K}$, where $D_k \in \mathbb{R}^{N_k\times t}$. In addition, K representative daily load profiles $D^C = \{\bar{D}_k\}_{k=1}^{K} = \{C_k\}_{k=1}^{K} \in \mathbb{R}^{K\times t}$ can be obtained by taking the average profile of each cluster.

C. Optimal Cluster Selection Stage

This stage aims to solve the challenges of **(C1)** and **(C2)** via 1) leveraging both pre-event and post-event baseline data of the tested dTOU customer to determine the optimal cluster for each day, 2) fully utilizing the synchronous daily baseline load patterns of the nonTOU customers obtained from the previous steps, 3) considering half-hour of the day as features to handle the problems of different event durations and start-times.

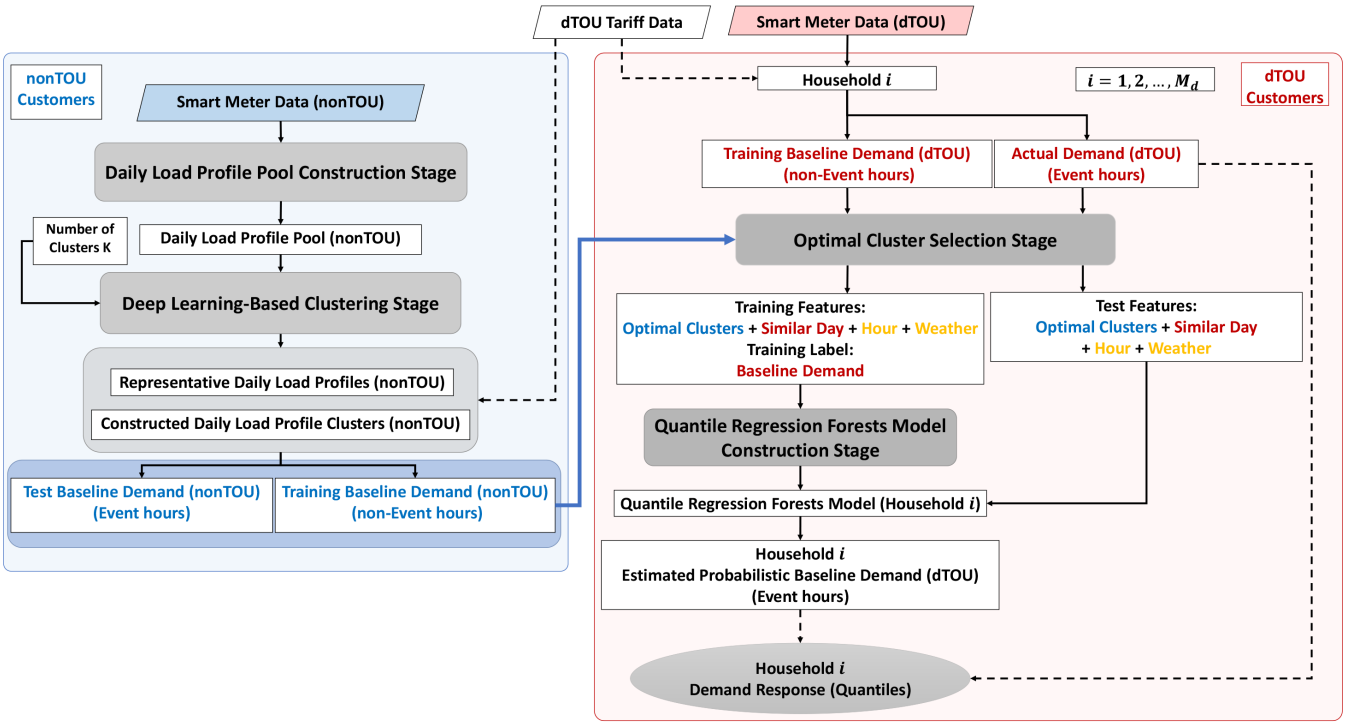

Fig. 1. The proposed probabilistic baseline estimation framework. Note that the blue color, the red color, and the yellow color indicate that the data are from the nonTOU customers, dTOU customers, and other relevant factors, respectively.

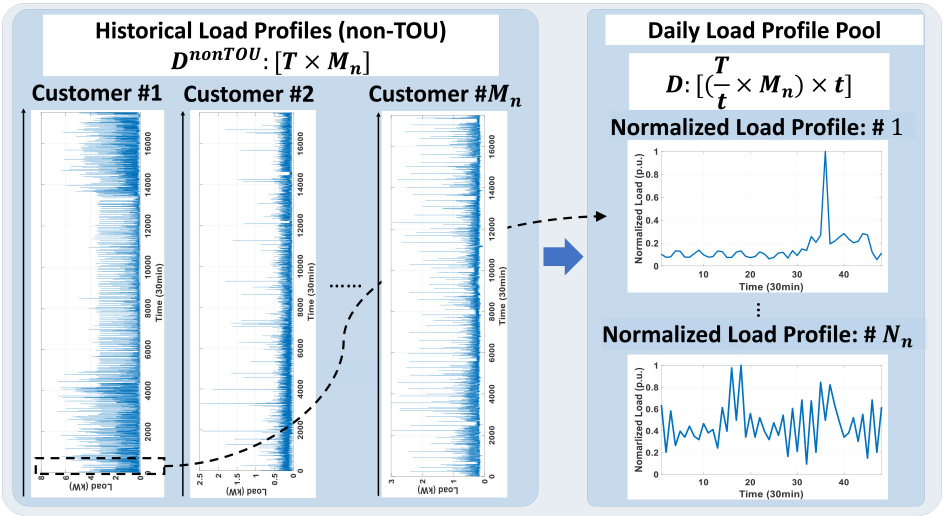

Fig. 2. The daily load profile pool construction stage.

For a given dTOU customer i, the features that will be used to construct the quantile regression forests model contain the most similar load patterns of the nonTOU customers for each day determined via optimal clusters selection, the average of the corresponding data over three most recent admissible days, and the temperature data, as shown in Fig. 1. Let Ω^E and Ω^B denote the set of days with and without events, respectively, where $|\Omega^E| + |\Omega^B| = T/t = 365$ and $\Omega = \Omega^E \cup \Omega^B$. An example of selecting the optimal cluster for each day $j = 1, ..., 365$ of customer i is illustrated in Fig. 3.

First, the j^{th} day's load data $d_{i,j}$ is normalised to $[0, 1]$ so that the most similar daily load patterns can be identified from the constructed Daily Load Profile Pool. Note that although we normalise the data in this step, the maximum value of $d_{i,j}$ will be saved and then used to transform the selected most similar patterns back to the original domain for retaining the magnitudes of $d_{i,j}$. Afterwards, according to the dTOU tariff data Π, if we have $j \in \Omega^E$, the load data of day j can be separated into the pre- and post-event data $d_{i,j}^{base} \in \mathbb{R}^{1 \times t_j^b}$ and the event data $d_{i,j}^{event} \in \mathbb{R}^{1 \times t_j^e}$, where t_j^b and t_j^e represent the non-event and event durations of day j, respectively. Otherwise, if $j \in \Omega^B$, $d_{i,j}^{base}$ will include all the data of day j

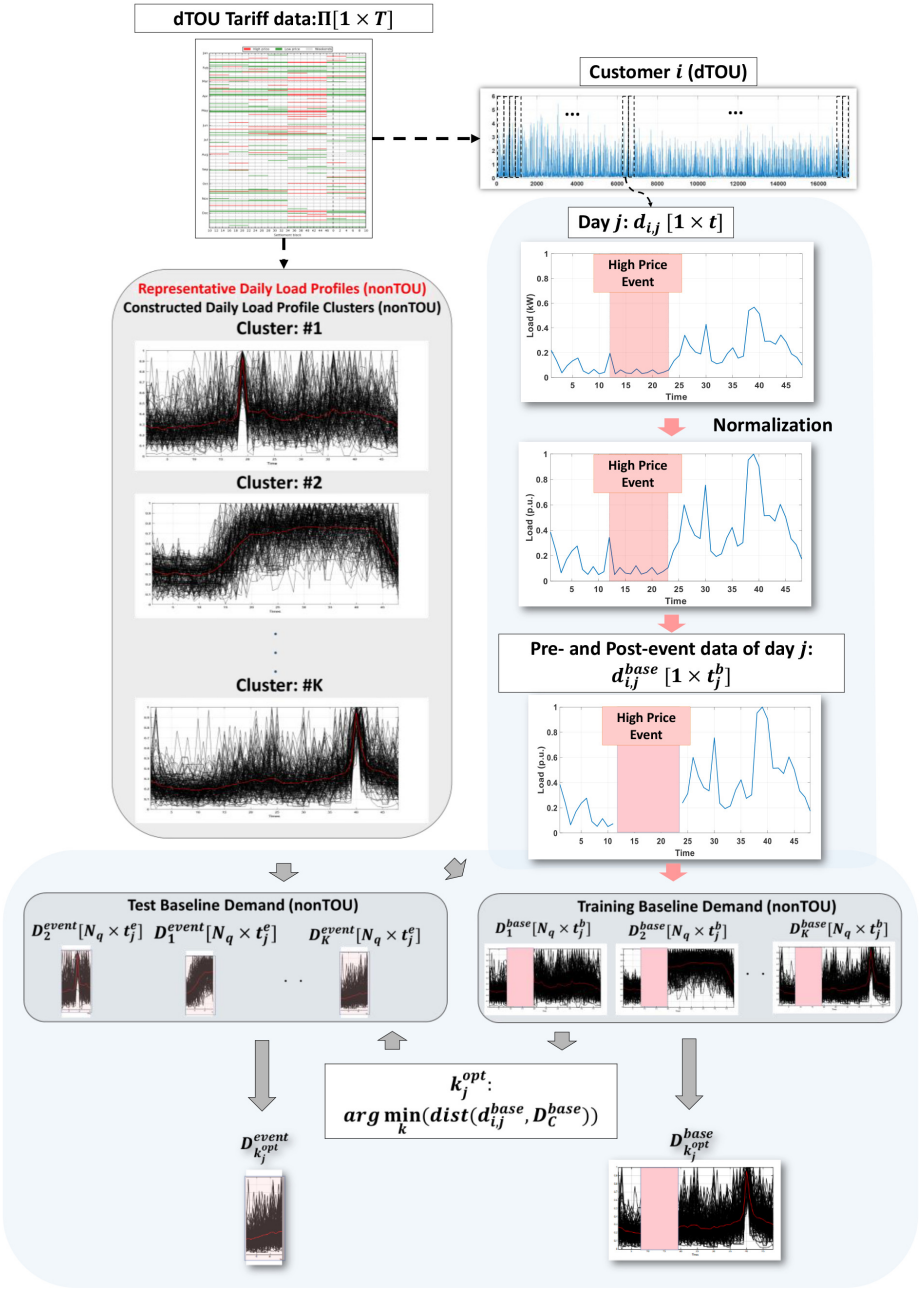

Fig. 3. The optimal cluster selection stage for day j of customer i.

(i.e., $t_j^b = t$ and $t_j^e = 0$). On the other hand, the K extracted representative daily load profiles as well as their corresponding clusters are also partitioned into event and non-event parts, denoted by the training baseline demand $C_k^{event} \in \mathbb{R}^{1\times t_j^e}$, $D_k^{event} \in \mathbb{R}^{N_k\times t_j^e}$and $C_k^{base} \in \mathbb{R}^{1\times t_j^b}$, $D_k^{base} \in \mathbb{R}^{N_k\times t_j^b}$, respectively, for $k = 1, ..., K$.

Subsequently, among the sets of training baseline demand $\{D_k^{base}\}_{k=1}^K$, the optimal cluster for day j is determined by selecting the $k_{i,j}^{opt}$ that leads to the minimum Euclidean distance between $d_{i,j}^{base}$ and C_k^{base}:

$$k_{i,j}^{opt} = \arg\min_k dist(d_{i,j}^{base}, C_k^{base}) \tag{4}$$

where

$$dist(d_{i,j}^{base}, C_k^{base}) = \sqrt{\sum_{h=1}^{t_j^b}(d_{i,j,h}^{base} - C_{k,h}^{base})^2} \tag{5}$$

After determine the $k_{i,j}^{opt}$, the optimal cluster with the most similar daily load patterns to day j of customer i can be obtained for training and testing (i.e., baseline estimation) purposes, defined as $D_{k_{i,j}^{opt}}^{base}$ and $D_{k_{i,j}^{opt}}^{event}$. Eventually, $D_{k_{i,j}^{opt}}^{base}$ and $D_{k_{i,j}^{opt}}^{event}$ are transformed back to the original kW domain by multiplying the pre-calculated maximum value of $d_{i,j}$. Combining the optimal clusters for all the days, the key part

of features used to build and test the regression model can be obtained as follows:

$$X_i^{Train\_OPT} = \{D_{k_{i,j}^{opt}}^{base}\}_{j\in\Omega}, X_i^{Test\_OPT} = \{D_{k_{i,j}^{opt}}^{event}\}_{j\in\Omega^E}. \tag{6}$$

It is important to note that all the $D_{k_{i,j}^{opt}}^{base}$ and $D_{k_{i,j}^{opt}}^{event}$ should have the same dimension as parts of input features, which can be achieved by taking the quantiles at the pth percentile where $p = [0.005, 0.010, ..., 0.995]$.

Beyond the features from the nonTOU customer, it is also crucial to exploit the information from the tested dTOU customer itself. As proposed in the literature, we consider the average of the corresponding data over three most recent admissible days as additional features, denoted by $X_i^{Train\_AVE}$ and $X_i^{Test\_AVE}$.

Finally, we consider the corresponding temperature data $X_i^{Train\_TEMP}$ and $X_i^{Test\_TEMP}$, and the index of hour $X_i^{Train\_HOUR}$ and $X_i^{Test\_HOUR}$ as additional features to further improve the performance and solve the challenges of **(C2)**. To sum up, the features X_i^{Train} used to construct the estimation model can be expressed as:

$$X_i^{Train} = [X_i^{Train\_OPT},\ X_i^{Train\_AVE},\ X_i^{Train\_TEMP},\ X_i^{Train\_HOUR}]. \tag{7}$$

Additionally, this stage also outputs the stack of target baselines during the non-event hours, defined in (8), which can be regarded as the response variable of the training explanatory variables X_i^{Train}.

$$Y_i^{Train} = \{d_{i,j}^{base} \times max(d_{i,j})\}_{j\in\Omega}, \tag{8}$$

D. Quantile Regression Forests Model Construction Stage

To solve the challenge of capturing the massive uncertainty of residential baseline demand **(C3)**, quantile regression forests model F_i is trained for each customer i based on the training explanatory variables X_i^{Train} and the training response variable Y_i^{Train}. As one of the most powerful probabilistic regression methods, the quantile regression forests method is a non-parametric method based on random forests that aims to estimate conditional quantiles for high-dimensional predictor variables [19]. Compared with a conventional random forest, the node of which retains only the mean of the samples, quantile regression forests consider the spread of the predictor (i.e., X_i^{Train}) by keeping the values of all samples in a node and therefore enabling the output of various estimation intervals, which can capture the uncertainties arising from various consumer behaviors. A detailed description of quantile regression forests is given in [19].

Given the input test features X_i^{Test}, the outputs of model F_i will be the estimated probabilistic baseline demand for customer i that includes all the dTOU tariff events. The output probabilistic estimated baseline $\hat{Y}_i^{Test}$ contain 99 quantiles at the pth percentile where $p = [0.01, 0, 02, ..., 0.99]$. Using equation (1), DR can be quantified for each customer by calculating the difference between the estimated baseline demand $\hat{Y}_i^{Test}$ and the actual measurements Y_i^{Test}. Note that, instead of using the real dTOU customers which cannot provide the real baseline load for evaluation during the dTOU events periods, we consider part of the nonTOU customers as the synthetic tested dTOU customers in the case study part so that the accuracy of the proposed framework can be assessed by calculating the evaluation metrics based on the estimated and actual baseline values, also represented by $\hat{Y}_i^{Test}$ and Y_i^{Test}, respectively.

IV. Deep Embedded Clustering

Unsupervised clustering for electrical customers is a crucial procedure for various applications in smart grid such as peak load estimation [20], load forecasting [21] and system operation and planning [22], [23]. Recently, the authors of [7] propose to calculate the baselines for residential loads based on clustering approaches. However, the conventional k-means method is used to classify customers into different groups based on their average daily load profiles, which may render it difficult to extract sufficient representative daily load patterns for all customers and therefore limit the baseline estimation performance, especially at the household level. To this end, neglecting information regarding customers and different load conditions (e.g., seasons and days of the week), the proposed clustering stage is completed based on the constructed daily load profile pool, which contains a considerable number of different patterns. For example, a pool of 2000 customers for 1 year of data consists of $2000 \times 365 = 730,000$ daily load profiles. Traditional distance-based clustering approaches (e.g., k-means, hierarchical clustering) use a distance measure to assess the similarity of data points and therefore, appropriate feature representation of the data is of great importance to enhance the clustering performance and obtain the "true" cluster label for each data point. This is because improved features can provide a better representative similarity matrix [24].

To this end, deep learning can be considered as a powerful tool that aims to learn high-level representations of raw data by using multiple layers of computational models [25], which has been successfully employed in image processing, speech recognition, objective detection, etc. Currently, the widespread deployment of smart meters in modern power systems provides valuable opportunities to understand electricity consumer behaviors by investigating the massive amount of fine-grained data. Consequently, deep learning approaches have become suitable for solving power system problems. The superior performance of deep learning has been demonstrated in various supervised learning based applications such as residential load forecasting [26], [18], system security assessment [27], socio-demographic information identification from smart meter data [28] and false data injection attacks detection [29]. However, very few work has been carried out regarding the topic of unsupervised learning, especially for clustering. A series of open questions have been raised regarding the model selection of deep neural networks (DNNs) and the definition of the clustering-oriented loss function [30].

Existing deep learning-based clustering consists of two-stage approaches and integrated approaches. An example of the two-stage approaches is presented in [31], which first

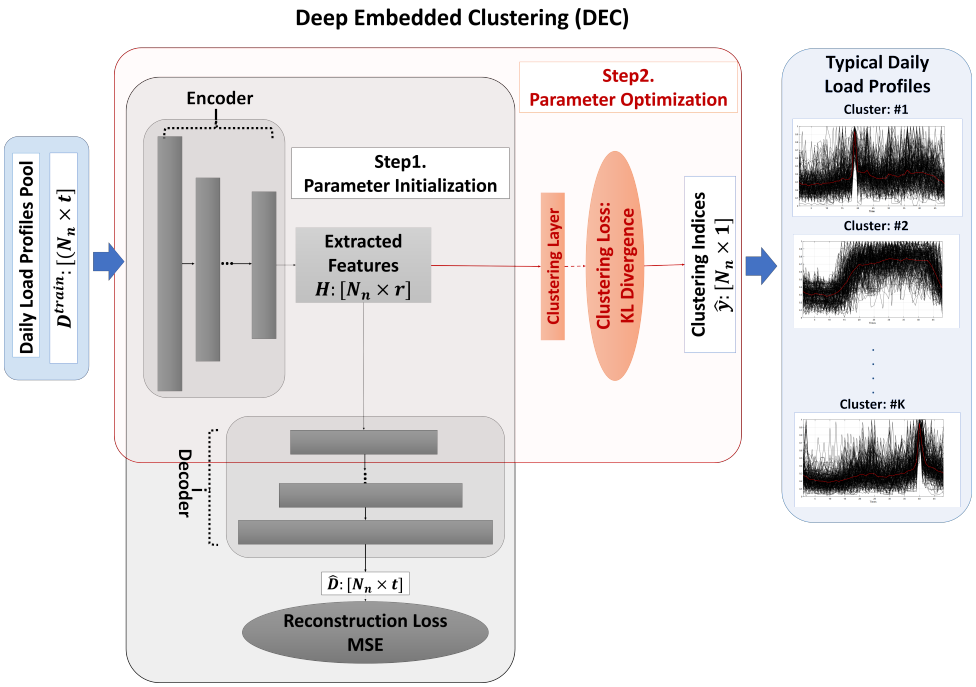

Fig. 4. The network structure of DEC. Fully connected layers are used to construct the encoder and decoder. Step 1 aims to initialize the parameters of DEC and then retain the trained encoder layers for Step 2, which further optimizes the parameters for minimizing the clustering loss.

employs an autoencoder to extract low-dimensional features and then uses the k-means method to cluster the data. The second type of deep learning clustering approaches is designed to combine the representation learning process and the clustering process into a single model while jointly minimizing the weighted sum of reconstruction loss and an explicitly defined clustering-oriented loss. Compared with the two-stage approaches, these integrated approaches can obtain effective features and cluster labels simultaneously. In this paper, we propose employing a novel integrated approach, namely, deep embedded clustering, which was proposed in [32], in the clustering stage of the proposed baseline estimation framework. As stated in [30], one of the most significant contributions of DEC is the proposition of the clustering loss, which works by employing highly confidential data points for supervision and then enabling these points to be distributed more densely for each cluster.

Let K and $D = D^{train} = [d_1, ..., d_{N_n}]^T \in \mathbb{R}^{N_n \times t}$ denote the target number of clusters and the *daily load profile pool* that need to be clustered, respectively. The constructed clusters and the corresponding centroids can be represented by $\{D^k\}_{k=1}^K$, where $D^k \in \mathbb{R}^{N_k \times t}$ and $D^C = [\bar{D}_1, ..., \bar{D}_K]^T \in \mathbb{R}^{K \times t}$. As illustrated in [32], DEC consists of network parameter initialization via training a deep autoencoder network and network parameter optimization for minimizing the Kullback-Leibler (KL) divergence-based clustering-oriented loss function. The overall workflow of the DEC method, presented in Fig. 4, is as follows.

A. Step 1. Parameter Initialization

This step is performed to initialize the proposed DEC network and retain the trained encoder layers for the next step. More specifically, DEC is first initialized with a deep autoencoder network (DAE) [33], which has been demonstrated as a powerful unsupervised learning model that can effectively extract informative and readily distinguishable representations in lower-dimensional space [33]. An autoencoder (AE) is an artificial neural network (ANN) that is trained to extract a hidden layer h to represent the input while minimizing the difference between the input original and output reconstructed datasets [34]. In particular, the AE network is composed of an encoder function $h = f_W(d) = s(Wd + b)$ and a decoder function $\hat{d} = g_{W^T}(h) = s(W^T h + b^T)$, where s is the nonlinear sigmoid activation function $s(h) = \frac{1}{1+\exp(-h)}$. For a deep autoencoder network, we define V and $N_V = [n_1, ..., n_V] \in \mathbb{R}^{V \times 1}$ (i.e., $n_1 = t$ and $n_V = r$) as the total number of layers and the number of neurons per layer, respectively. Given the input data $D = [d_1, ..., d_{N_n}]^T \in \mathbb{R}^{N_n \times t}$, the DAE network is initialized layer by layer, with each layer being a denoising autoencoder with random corruption, defined as follows:

$$\tilde{h}_v \leftarrow DP(h_v) \tag{9}$$

$$h_{v+1} = s(W_v \tilde{h}_v + b_v), \forall v \in [1, ..., V] \tag{10}$$

where DP represents the $Dropout(\cdot)$ function, which is a stochastic mapping function used to randomly set a number of the input dimensions to zero, as defined in [35]. The denoising DAE is trained to minimize the reconstruction error J between d and $\hat{d}$ as follows:

$$J = \|\hat{d} - d\|_2^2 \tag{11}$$

Greedy layerwise pretraining is employed to initialize the autoencoder network by training V blocks of denoising autoencoders, with the output layer of each block being used as

the input layer of the next one. Afterwards, the deep encoder and decoder network is unfolded and initialized. Finally, global fine-tuning is performed to further optimize the parameters via backpropagation by minimizing the reconstruction loss. As described in [32], the decoder layers are discarded, and the encoder layers are used as the initial feature extractor, as shown in the orange rectangle area (i.e., Step 2) of Fig. 4.

B. Step 2. Parameter Optimization

The target of this step is to further optimize the parameters of deep encoder layers and update the cluster centroids through backpropagation until the convergence criterion is achieved. Based on the extracted features $H = [h_1, ..., h_{N_n}]$ in the previous step, conventional k-means clustering is performed to obtain K initial cluster centroids $D^C = [\bar{D}_1, ..., \bar{D}_K]^T = [\mu_1, ..., \mu_K]^T \in \mathbb{R}^{K \times t}$. In Step 2, we first need to compute the similarity $q_{i,j}$ between each embedded point h_i, where $i \in [1, ..., N_n]$, and each centroid μ_k, where $k \in [1, ..., K]$, measured based on Student's t-distribution as follows:

$$q_{i,k} = \frac{(1 + \|h_i - \mu_k\|^2)^{-1}}{\sum_k (1 + \|h_i - \mu_k\|^2)^{-1}} \tag{12}$$

Note that $q_{i,k}$ can be considered as the probability of assigning sample i to cluster k. Then the retained encoder layers are fine-tuned by minimizing the KL divergence between q_i and the auxiliary distribution p_i, defined as follows:

$$L = KL(P\|Q) = \sum_i \sum_j p_{i,k} \log \frac{p_{i,k}}{q_{i,k}} \tag{13}$$

where $p_{i,k}$ is expressed as:

$$p_{i,k} = \frac{q_{i,k}^2 / \sum_i q_{i,k}}{\sum_i (q_{i,k}^2 / \sum_i q_{i,k})} \tag{14}$$

It is important to note that minimizing $KL(P\|Q)$ is a form of self-training [30]. Afterwards, for each sample d_i, the cluster label y_i is re-signed by finding the optimal index k^* that contains the maximum value of all $q_{i,k}$ for $k = 1, ..., K$ (i.e., $y_i = arg\max_k q_{i,k}$). Regarding the parameter optimization procedure of the cluster centroids $[\mu_1, ..., \mu_K]^T \in \mathbb{R}^{K \times t}$ and the parameters of the deep encoder layers, stochastic gradient descent (SGD) is used to calculate the gradients of L with respect to each h_i and each cluster centroid μ_k as follows:

$$\frac{\partial L}{\partial h_i} = 2 \times \sum_k (1 + 1 + \|h_i - \mu_k\|^2)^{-1} \times (p_{i,k} - q_{i,k})(h_i - \mu_k) \tag{15}$$

$$\frac{\partial L}{\partial \mu_k} = -2 \times \sum_i (1 + 1 + \|h_i - \mu_k\|^2)^{-1} \times (p_{i,k} - q_{i,k})(h_i - \mu_k) \tag{16}$$

Then, new $\frac{\partial L}{\partial h_i}$ and $\frac{\partial L}{\partial \mu_k}$ are passed down to update the retained deep encoder layer parameters and the cluster centroids, respectively. Note that other optimizers such as Adam, Adamax, Adadelta and AdaGrad can also be used as alternatives for parameter optimization. The abovementioned procedures are iteratively carried out until the predefined convergence criterion ϵ is achieved. In other words, DEC stops when less than $\epsilon\%$ labels are changed between every two iterations. The whole DEC method is outlined in Algorithm 1.

Algorithm 1 Deep Embedded Clustering

Input: Daily load profile pool: $D = [d_1, ..., d_{N_n}]^T \in \mathbb{R}^{N_n \times t}$; Number of clusters: $K \in \mathbb{Z}_{>0}$; Maximum number of iterations: $\overline{iter}$; Convergence criterion: ϵ; Target distribution update interval: I_U.

Output: Cluster labels: $\hat{y} \in \mathbb{Z}_{>0}^{N_n \times 1}$

***Step 1**: Initialize the parameters of DEC by constructing and training a deep denoising autoencoder network with V layers, with $N_V = [n_1, ..., n_V]$ neurons for each layer. Retain the trained encoder layers f_W for the next step.*

***Step 2**: Initialize cluster centroids $[\mu_1, ..., \mu_K]^T \in \mathbb{R}^{K \times t}$ via k-means method based on the extracted features $H = [h_1, ..., h_{N_n}] = f_W(D) \in \mathbb{R}^{N_n \times r}$.*

$\hat{y} = Kmeans(H, K)$.
for $iter$ = 1:$\overline{iter}$ **do**
  if $Modulo(iter, I_U) == 0$ **then**
    Transform data from $D = [d_1, ..., d_{N_n}]^T$ to the feature domain $H = [h_1, ..., h_{N_n}]^T$.
    for i = 1:N_n **do**
      $h_i = f_W(d_i)$
    end for
    Update P using (12) and (14).
    $\hat{y}^0 = \hat{y}$
    Update cluster labels.
    for i = 1:N_n **do**
      $\hat{y}_i = arg\max_k(q_{i,k})$
    end for
    if $\frac{\#(\hat{y}^0 \neq \hat{y})}{N_n} < \epsilon$ **then**
      Stop
    end if
  end if
  Pick a batch of data $D^S \subset D$.
  Optimize the parameters of deep encoder layers W and b, and update the cluster centroids $[\mu_1, ..., \mu_K]^T$ based on D^S via backpropagation based on (15) and (16).
end for

V. Case Studies

A. Data Description

The smart meter data used in this paper is collected in the Low Carbon London program, which is a technology demonstrator supported by customers via the Low Carbon Network Fund (LCNF) [16], [36]. Specifically, the LCL demand dataset consists of 17,520 half-hourly measurements of demand across 5,112 customers in kW for a full calendar year from 1^{st} January 2013 to 31^{st} December 2013. In this program, the residential dTOU tariff was trialed to investigate and quantify electricity customers' attitudes and responsiveness to time-varying electricity pricing. All 5,112 households in the trial were categorized into either the non-time-of-use (nonTOU) group (i.e., $M_n = 4,068$ households) or the dTOU group (i.e, $M_d = 1,044$ households), which receive a standard flat tariff (i.e., 14.228 pence/kWh) and the experimental dTOU tariff (i.e., a high price of 67.2 pence/kWh, a default price of 11.76 pence/kWh, and a low price of 3.99 pence/kWh), respectively. In particular, the events designed for the trial

TABLE I
A SUMMARY OF THE SIMULATION SCENARIOS

Scenario ID	Replications #	Training/Test Customers #	Event Type	Event Duration	Methods	Others
S1	10	500/50	Both	All	M1-M5	Different numbers of clusters (M5)
S2				3-hour	M5	-
S3	10	500/50	High-price	6-hour	M5	-
S4				12-hour	M5	-
S5				3-hour	M5	-
S6	10	500/50	Low-price	6-hour	M5	-
S7				12-hour	M5	-
S8	100	50/5	Both	All	M1, M3, M5	T-test (M1, M3, M5)
S9					M5	Robustness test (bad data) (M5)

consist of constraint management (CM) events, which are aimed at managing the distribution network constraints, and supply following (SF) events, which are proposed for supply demand balancing. In particular, the SF events are designed to provide a general insight into consumer response to dTOU tariffs. Therefore, in this work, we focus on the SF events determined by a randomized block design, which include 45 high-price events and 48 low-price events in total with 3-hour, 6-hour, and 12-hour durations randomly scattered throughout the year of the trial. More detailed information regarding the trial design and the tested dataset can be found in [37].

TABLE II
PARAMETER SETTING FOR DEC

Parameter	Value
Number of layers	5
Number of neurons	48-60-60-800-20
Layer type	Fully connected
Batch size	1000
Convergence threshold	0.1%
Optimizer	Adam
Learning rate	0.001
Maximum number of iterations	2e5
Distribution update interval	2000

B. Methods for Comparison and Simulation Scenarios

A series of tested methods are implemented in this paper for comparison in terms of the point estimation performance and the probabilistic estimation performance:

1) ***M1-Simple Average***: Estimate the baseline demand during the event period of the test day by averaging the corresponding data over the three most recent admissible days [8];

2) ***M2-Average Daily Load Profile + k-Means Clustering + Quantile Regression Forests***: As illustrated in [7], the baseline demand during the event period is proposed to correspond to the clusters, which are constructed based on the set of average daily load profiles for each customer. To demonstrate the importance of the proposed *Daily Load Profile Pool Construction Stage*, we consider the average daily load profiles of each customer to estimate the baseline, following the steps presented in Section III;

3) ***M3-Average Daily Load Profile + k-Means Clustering + Gaussian Process***: In [14], the Gaussian process is demonstrated to be an appropriate model for probabilistic baseline estimation. In this case, we use the Gaussian process to predict the baseline demand after performing the clustering procedure instead of using quantile regression forests;

4) ***M4-Proposed + k-Means Clustering + Quantile Regression Forests***: Based on the proposed framework, the k-means technique is employed in this tested method to demonstrate the superior performance of the proposed daily load profile pool and the DEC method;

5) ***M5-Proposed + DEC + Quantile Regression Forests***: The proposed probabilistic baseline estimation framework. Note that the parameter setting for DEC in the following simulations is given in Table II.

Additionally, all the simulation scenarios (i.e., S1-S9) carried out in this paper are summarized in Table I. In the following parts, the results of scenario S1 are shown in Section V-C, Section V-D1 and Section V-D2. Scenarios S2-S7 are conducted in Section V-D3. Furthermore, Scenarios S8 and S9 are investigated in Section V-E and Section V-F, respectively.

C. Deterministic Estimation Performance Analysis

To quantitatively evaluate the performance of the baseline estimation methods, the experiments are conducted on ten sets of *training* (500 customers) and *test* (50 customers) customers selected from the non-dTOU group. Note that the real event periods of dTOU tariffs correspond to the customers in *test sets* for a full calendar year from 1^{st} January 2013 to 31^{st} December 2013. In this way, the actual demand of the test dataset can be regarded as the benchmark baseline and therefore can be used to compute the estimation accuracy in terms of the considered evaluation metrics. Among the five tested methods, all the methods (i.e., M1-M5) are compared in terms of their point estimation results, whereas the probabilistic estimated baselines are evaluated for M2-M5. First, the *mean squared error (MSE)* and the *average relative error (ARE)* [7] are employed to assess the deterministic estimation results for M1-M5 regarding the accuracy and bias, receptively, as shown in Table III, using the average metric values over the ten test sets.

The superior performance of the proposed method M5 can be demonstrated in terms of the point estimation accuracy, with it having the lowest MSE value, achieving approximately 13.25% improvements when compared with the conventional simple average method M1. In addition, M4 exhibits an approximately 6.40% lower MSE value than that of M2, which highlights the importance and effectiveness of the proposed daily load profile pool. For M2, it can be concluded that quantile regression can lead to more accurate estimation results than those obtained using the Gaussian process, exhibiting the highest MSE value among all the tested methods. On the other

hand, the estimation bias, quantified via the ARE, presents slightly different results; M1 has the lowest ARE value, which indicates that it has the best performance, whereas M3 has negative values, with its largest ARE value being -0.0273. It is important to note that M1 is a deterministic estimation method, and among the tested probabilistic approaches, the proposed framework M5 with DEC achieves roughly 86.75% and 60.78% enhancements in terms of the bias when compared with M2 and M4, respectively.

TABLE III
DETERMINISTIC BASELINE ESTIMATION PERFORMANCE

	M1	M2	M3	M4	M5
MSE (kW^2)	0.0302	0.0297	0.2592	0.0278	**0.0260**
ARE (kW)	**0.0009**	0.0151	-0.0273	0.0051	0.0020

D. Probabilistic Estimation Performance Analysis

*1) **Methods Comparison**:* Beyond the analysis of the point estimation results, the main target of this work is to characterize uncertainties by performing probabilistic baseline estimation. Calibration, reliability, and sharpness are three main factors that indicate the consistency, the variation, and the tightness of the estimated distribution. As two of the most widely used comprehensive evaluation metrics for probabilistic estimation or forecasting, the *pinball loss (PL)* and *Winkler score (WS)* are employed in this work. Additionally, we use the *prediction interval coverage probability (PICP)* as another metric to measure the probability that the actual baseline demand falls into the upper and lower bounds of the estimated probabilistic baseline demand. Note that detailed formulations of the aforementioned metrics can be found in [17].

Table IV lists the calculated PLs, Winkler scores and PICP values for all the examined methods. As shown below, the order of the probabilistic estimations performance is highly consistent with the results shown in Table III in terms of all three employed evaluation metrics. Specifically, the introduction of the proposed daily load profile pool can effectively decrease the PL and Winkler score values from 0.0351 (M2) to 0.0330 (M4) and from 8.2426 (M2) to 8.0536 (M4), respectively.

TABLE IV
PROBABILISTIC BASELINE ESTIMATION PERFORMANCE

	M2	M3	M4	M5
PL(kW)	0.0351	0.1575	0.0330	**0.0315**
Winkler Score (kW)	8.2426	45.5724	8.0536	**7.8784**
PICP	0.8050	0.4626	0.9252	**0.9282**

Furthermore, a better prediction interval coverage probability can be obtained by performing clustering based on the pool instead of the average load profiles for individual customers, as indicated by the approximately 12% higher PICP value of M4 than that of M2. Finally, the results of M5 demonstrate that the proposed DEC clustering method can further improve the probabilistic baseline estimation performance, as evidenced by the approximately 3.64% lower PL, 2.18% lower Winkler score, and 0.32% higher PICP. Overall, the improvements from the existing probabilistic baseline estimation method M3 to the proposed method M5 can achieve about 80%, 82.71%, and 50% regarding the evaluation metric values of PL, Winkler score, and PICP, respectively, which can effectively demonstrate the significant contribution of the proposed method.

*2) **Different Numbers of Clusters**:* Regarding the selection of the optimal number of clusters K for the proposed method M5, unlike the conventional clustering validation indicator-based approach, the calculated evaluation metrics for the probabilistic estimation are directly used to determine the most appropriate K while jointly considering the main factors of calibration, reliability and sharpness based on the PL, Winkler score and PICP.

In this case, we use M5 with different numbers of clusters $K = [10, 20, 30, 100, 300, 500]$ based on the constructed *daily load profile pool*; the corresponding evaluation metrics values are given in Fig. 5. In terms of the PICP, it can be observed that K = 300 yields the highest coverage probabilities. Nevertheless, it is not sufficient to take into account only the probability of estimation interval coverage. This is due to the fact that a larger interval will result in better coverage but may result in a significantly inaccurate estimated baseline when compared with the actual one. On the other hand, as K increases, both the PL and WS values tend to gradually grow after $K = 20$, which can be considered as the optimal number of clusters with the lowest PL and WS values in this case. Regarding the computational cost, for $K = 20$, the DEC clustering method ran $31,860$ iterations to achieve the convergence criterion and the whole procedure takes approximately $1,929.36$ seconds.

*3) **Different Durations of Events**:* In this trial, the SF events consist of high-price events and low-price events with 3-hour, 6-hour, and 12-hour durations. Table V presents the probabilistic baseline estimation results for different event durations of high price and low price. Interestingly, the results illustrate that there is no evident relationship between the evaluation metric values and the duration of events, indicated by the variational trends of the PL, WS, and PICP values from 3-hour events to 12-hour events. This finding demonstrates the robustness and effectiveness of the proposed framework when addressing various event durations. In particular, the regression model in the proposed framework is constructed based on the non-event data for the whole year instead of considering only the pre- and post-event hours, thus improving the volume of data used to build the regression model, which can make the method suitable for estimating the baseline for longer durations, such as 24-hour events.

TABLE V
PROBABILISTIC BASELINE ESTIMATION PERFORMANCE FOR DIFFERENT EVENT DURATIONS

	High-Price Event			Low-Price Event		
	3 h	6 h	12 h	3 h	6 h	12 h
PL	0.0236	0.0319	0.0288	0.0254	0.0386	0.0325
WS	2.6933	5.5455	11.2016	4.3263	5.3081	12.7753
PICP	0.9640	0.9181	0.9317	0.8846	0.9220	0.9181

Moreover, Fig. 6 illustrates the actual baseline and estimated baseline obtained by the proposed deep-learning-based method M5 based on a series of random days selected from one

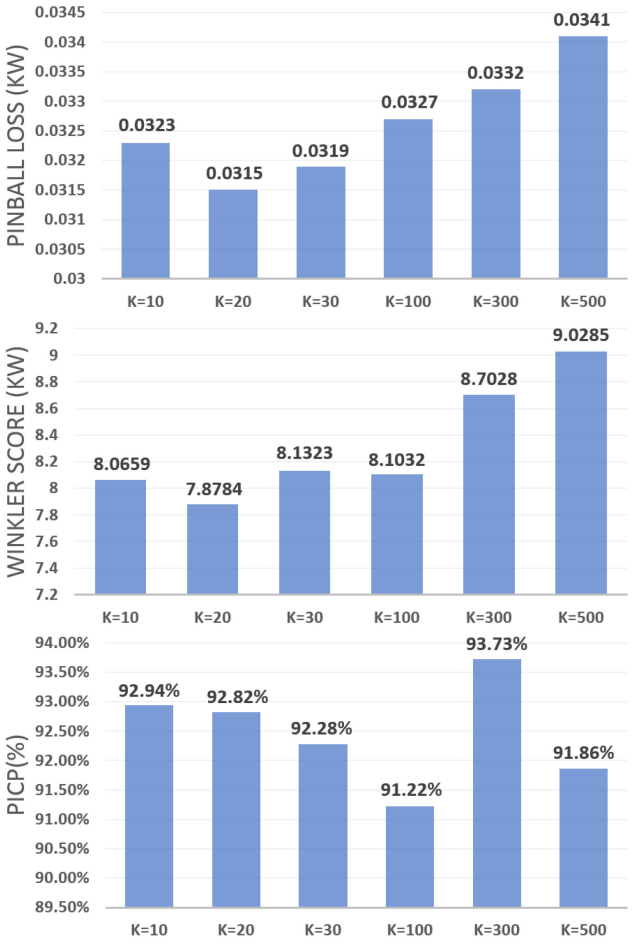

Fig. 5. Probabilistic estimation results for different numbers of clusters.

customer's whole year load profile for different dTOU prices (i.e., high price and low price), event durations (e.g., 3 hours and 6 hours), and start times (e.g., 2:00am, 5:00am, 14:00pm and 23:00pm). Note that detailed information regarding the selected days and events is shown in the tile of each subplot. The upper bound, lower bound and mean profile of the estimated probabilistic baseline demand are represented by red dashed lines and a red solid line, respectively. As shown in the figure, relatively lower electricity consumption occurring at midnight and in the morning hours can be well estimated for both 3-hour (Day 10) and 6-hour (Day 21) low-price events as well as 3-hour high-price events (Day 73, Day 101). During the daytime, more fluctuations in the baseline demand can be observed during the event periods of Day 25 and Day 304. These fluctuations increase the difficulty in capturing the series of sudden peaks. For Day 170, which includes 3-hour high-price events starting from 14:00pm, the results show larger estimation intervals but trends similar to those of the actual baselines. Consequently, such visual inspection, combined with the evaluation results shown in Tables III and IV, consistently demonstrates that the combination of the daily load profile pool and DEC (i.e., M5) can be regarded as a reliable and effective method for estimating the baseline demand considering the uncertainty for different types of events.

E. Performance Comparison via Statistical Test

In order to highlight the statistically significantly superior performance of the proposed approach, an additional case study based on T-test (two-sample, one-tailed) is conducted. In this case, 100 replications (randomly selected 100 training and 5 test customers) are performed and comparisons are made based on mean performance. It is notable that we make the comparisons between the proposed approach M5 and the existing approaches M1 and M3 (i.e., M2 and M4 are also originally developed in this paper) to present the statistically significant results regarding the deterministic and probabilistic estimation performance based on the MSE and the PL. The average value and the variance of the MSE and the PL for M1, M3 and M5 are presented in Table VI. As can be seen, the proposed baselines estimation method M5 also exhibits lower mean and variance values than the existing approaches M1 and M3. More specifically, compared with M1, the proposed method M5 exhibits approximately 37.54% lower mean and 62.07% lower variance of the calculated MSE values, respectively. On the other hand, regarding the probabilistic estimation performance, M5 presents about 67.04% and 60.00% improvements in the calculated PL values in terms of their mean and variance when compared with M3.

TABLE VI
MEAN AND VARIANCE OF THE CALCULATED METRIC VALUES (100 REPLICATIONS)

	MSE(mean)	MSE(var)	PL(mean)	PL(var)
M1	0.0618	0.0058	-	-
M3	0.2733	0.0487	0.1065	0.0015
M5	0.0386	0.0022	0.0351	0.0006

Regarding the T-test, the null hypothesis is the sample mean of the evaluation metric value for M1 (or M3) is greater than or equal to the sample mean of the evaluation metric value for the proposed M5 at the 5% significance level. Furthermore, the results of the T-tests are shown in Table VII with the p-values. Overall, the results illustrate that the all the conducted T-tests do not reject the null hypotheses at the 5% significance level and thus demonstrating the statistically significantly improved performance of the proposed estimation approach.

TABLE VII
RESULTS OF THE TWO-SAMPLE, ONE-TAILED T-TESTS FOR DIFFERENT EVALUATION METRICS (100 REPLICATIONS)

	M1>M5	M3>M5
MSE (test result/p-value)	not reject/p=0.9948	not reject/p=1
PL(test result/p-value)	-	not reject/p=1

F. Robustness Test with Bad Data

The aim of the final case study lies in showing the effect of bad data on baseline estimation and demonstrates the robustness of the proposed method. Note that the examined bad data is constructed by adding different levels of Gaussian noise array to the original LCL dataset with mean equal to 0 and

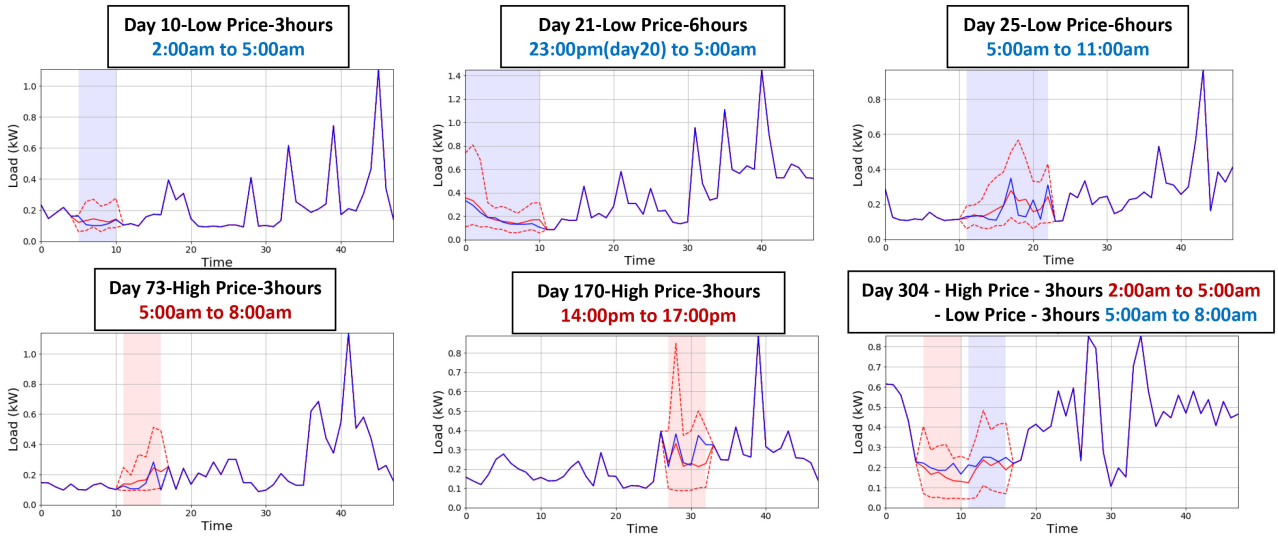

Fig. 6. Actual and estimated baseline demand obtained via the proposed method M5 based on a series of random days with different types and durations of events (blue area: low-price event; red area: high-price event). Note that the upper/lower bounds and the mean profile are indicated by two red dashed lines and one red solid line, respectively.

standard deviation (noise level) equal to $[0, 0.01, 0.05, 0.1]$, respectively. Note that the maximum standard deviation (i.e. 0.1) is determined according to the load magnitudes of the test customers.

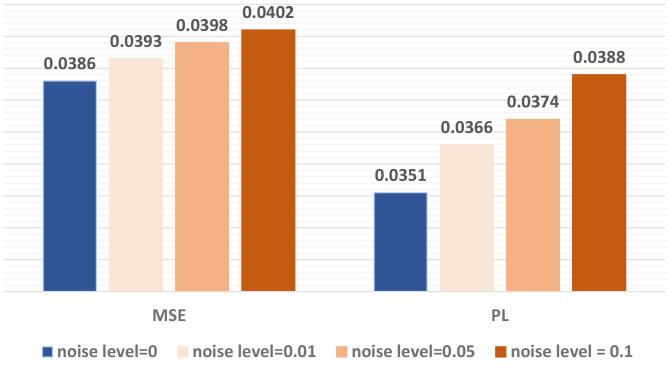

Fig. 7. Estimation results across different levels of noise (M5)

.

Under the simulation scenario S9, the estimation results of the proposed method M5 are given in Fig. 7 with different levels of bad data (noisy data). First, it can be observed that the calculated MSE and PL values are nearly at the same level for the cases of clean data (noise level $= 0$) and slightly noisy data (noise level $= 0.01$). Beyond that, it can be concluded that the increasing level of injected noise indeed reduces the estimability of baseline demand indicated by the increased MSE and PL values. Nevertheless, under the worst case (i.e., noise level $= 0.1$, which is 10 times larger than 0.01), the calculated average MSE and average PL values are only approximately 1.02 times and 1.11 times larger than the case of noise level equal to 0.01, respectively. Additionally, although the metric values get increased, compared with the results of M1 and M3 shown in Table VI, M5 with noisy data still outperforms conventional approaches with clean data. To this end, the above results demonstrate that the proposed method M5 can exhibit robust performance even in the context of bad data.

VI. Conclusions

This paper proposes a novel probabilistic baseline estimation framework which integrates the deep embedded clustering to capture the uncertainties of the baseline demand at the household level. In particular, the *Daily Load Profile Pool Construction Stage* and the *Deep Learning-Based Clustering Stage* effectively extract the information from nonTOU customers by identifying representative daily baseline load patterns from a pool, consisting of a massive number of possible daily baseline shapes. Then based on the pre-event and post-event baseline load of the tested dTOU customer, the *Optimal Cluster Selection Stage* aims to construct the training and testing features for the *Quantile Regression Forests Model Construction Stage* to build a quantile regression forests model for each customer, which can capture the massive uncertainty of residential baseline demand. With the features of hour of the day, different event durations and start-times can be handled properly. Case studies are performed based on real smart meter data corresponding to the nonTOU group of the LCL project. The superior performance of the proposed framework is demonstrated, with the estimation results gradually improving from the simple average approach to the clustering-based approach, from the customer's average daily profiles to the daily load profile pool, and from the conventional clustering method to the advanced deep learning-based method. Compared with the conventional baseline estimation approaches (i.e., deterministic: M1, probabilistic: M3), the proposed method M5 shows approximately 37.54% and 67.04% improvements regarding the mean values of MSE and PL, respectively. Additionally, the results of T-test demonstrate the statistically significantly improved performance of the proposed estimation approach.

In future work, the proposed model will be further enhanced regarding its robustness to bad data. Also, we would like to explore more advanced estimation model (e.g., deep neural network) to capture the forward and backward dependencies in times series data. Moreover, we can investigate the real impact of the proposed baseline estimation method in cost saving to demonstrate its practical benefits. Additionally, we would like to implement the proposed method for system level baseline demand estimation to quantify the DR at aggregated level.

REFERENCES

[1] S. Nan, M. Zhou, and G. Li, "Optimal residential community demand response scheduling in smart grid," *Applied Energy*, vol. 210, pp. 1280–1289, Jan. 2018.

[2] A. Asadinejad and K. Tomsovic, "Optimal use of incentive and price based demand response to reduce costs and price volatility," *Electric Power Systems Research*, vol. 144, pp. 215–223, Mar. 2017.

[3] K. McKenna and A. Keane, "Residential load modeling of price-based demand response for network impact studies," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2285–2294, Sept. 2016.

[4] V. Azarova, D. Engel, C. Ferner, A. Kollmann, and J. Reichl, "Exploring the impact of network tariffs on household electricity expenditures using load profiles and socio-economic characteristics," *Nature Energy*, vol. 3, pp. 317–325, Mar. 2018.

[5] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, in press.

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

[7] Y. Zhang, W. Chen, R. Xu, and J. Black, "A cluster-based method for calculating baselines for residential loads," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2368–2377, Sep. 2016.

[8] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, "Statistical analysis of baseline load models for non-residential buildings," *Energy and Buildings*, vol. 41, no. 4, pp. 374–381, Apr. 2009.

[9] Y. Wi, J. Kim, S. Joo, J. Park, and J. Oh, "Customer baseline load (cbl) calculation using exponential smoothing model with weather adjustment," in *2009 Transmission Distribution Conference Exposition: Asia and Pacific*, Oct 2009, pp. 1–4.

[10] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, "Estimating demand response load impacts: Evaluation of baselineload models for non-residential buildings in california," Ernest Orlando Lawrence Berkeley NationalLaboratory, Berkeley, CA (US), Tech. Rep., 2008.

[11] L. Hatton, P. Charpentier, and E. Matzner-Lber, "Statistical estimation of the residential baseline," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1752–1759, May 2016.

[12] S. Park, S. Ryu, Y. Choi, J. Kim, and H. Kim, "Data-driven baseline estimation of residential buildings for demand response," *Energies*, vol. 8, no. 9, pp. 10 239–10 259, Sept. 2015.

[13] Y. Weng, J. Yu, and R. Rajagopal, "Probabilistic baseline estimation based on load patterns for better residential customer rewards," *International Journal of Electrical Power Energy Systems*, vol. 100, pp. 508–516, Sept. 2018.

[14] Y. Weng and R. Rajagopal, "Probabilistic baseline estimation via gaussian process," in *2015 IEEE Power Energy Society General Meeting*, July 2015, pp. 1–5.

[15] M. Valls, A. Bello, J. Reneses, and P. Fras, "Probabilistic characterization of electricity consumer responsiveness to economic incentives," *Applied Energy*, vol. 216, pp. 296–310, Apr. 2018.

[16] J. Schofield, R. Carmichael, S. Tindemans, M. Woolf, M. Bilton, and G. Strbac, "Residential consumer responsiveness to time-varying pricing," Imperial College London, Tech. Rep., 2014.

[17] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, Jul.-Sept. 2016.

[18] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-term residential load forecasting based on resident behaviour learning," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1087–1088, Jan. 2018.

[19] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, pp. 983–999, Jun. 2006.

[20] M. Sun, Y. Wang, G. Strbac, and C. Kang, "Probabilistic peak load estimation in smart cities using smart meter data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1608–1618, Feb. 2019.

[21] Y. Wang, Q. Chen, M. Sun, C. Kang, and Q. Xia, "An ensemble forecasting method for the aggregated load with subprofiles," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3906–3908, Jul. 2018.

[22] M. Sun, F. Teng, I. Konstantelos, and G. Strbac, "An objective-based scenario selection method for transmission network expansion planning with multivariate stochasticity in load and renewable energy sources," *Energy*, vol. 145, pp. 871–885, Feb. 2018.

[23] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, Jun. 2012.

[24] E. Gultepe and M. Makrehchi, "Improving clustering performance using independent component analysis and unsupervised feature learning," *Human-centric Computing and Information Sciences*, vol. 8, no. 1, p. 25, Aug. 2018.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[26] H. Shi, M. Xu, and R. Li, "Deep learning for household load forecastinga novel pooling deep rnn," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5271–5280, Sep. 2018.

[27] M. Sun, I. Konstantelos, and G. Strbac, "A deep learning-based feature extraction framework for system security assessment," *IEEE Transactions on Smart Grid*, in press.

[28] Y. Wang, Q. Chen, D. Gan, J. Yang, D. S. Kirschen, and C. Kang, "Deep learning-based socio-demographic information identification from smart meter data," *IEEE Transactions on Smart Grid*, in press.

[29] Y. He, G. J. Mendis, and J. Wei, "Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2505–2516, Sep. 2017.

[30] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 1753–1759.

[31] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu, "Learning deep representations for graph clustering," in *AAAI*, 2014, pp. 1293–1299.

[32] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, 2016, pp. 478–487.

[33] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[34] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jun. 2014.

[36] M. Sun, I. Konstantelos, and G. Strbac, "C-vine copula mixture model for clustering of residential electrical load pattern data," *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 2382–2393, May 2017.

[37] J. Schofield, S. Tindemans, R. Carmichael, M. Woolf, M. Bilton, and G. Strbac, "Low carbon london project: Data from the dynamic time-of-use electricity pricing trial, 2013," Tech. Rep., Jan. 2016.