IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2020, accepted August 21, 2020, date of publication August 27, 2020, date of current version September 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019807

Modeling and Analysis of Residential Electricity Consumption Statistics: A Tracy-Widom Mixture Density Approximation

SHIWEN LIAO, (Student Member, IEEE), LU WEI, (Member, IEEE), TAEHYUNG KIM, (Senior Member, IEEE), AND WENCONG SU, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, University of Michigan-Dearborn, Dearborn, MI 48128, USA
Corresponding authors: Lu Wei (luwe@umich.edu) and Wencong Su (wencong@umich.edu)

ABSTRACT Residential electricity loads highly fluctuate which makes it challenging to predict and estimate changes and anomalies. This article studies the statistical distribution of the maximum load of residential settlements and proposes a new model that is more reliable than conventional methods to predict extreme events in residential electricity consumption. Specially, a multimodal Tracy-Widom distribution is proposed to characterize the maximum residential electricity consumption data. We also propose a numerical method to approximate the density function of the multimodal Tracy-Widom distribution as opposed to the conventional approach in the literature, where a multimodal normal distribution is utilized. A simulated electricity consumption data set is used to test and validate the proposed methods. The results demonstrate that the multimodal Tracy-Widom distribution is more accurate than conventional methods in modeling residential electricity consumption data. In addition, the numerical results show that residential electricity consumption behavior is determined by certain socioeconomic factors for which correlations with household electricity consumption exist.

INDEX TERMS Power system, residential electricity load, electricity consumption factor, peak load density, extreme value distribution, Tracy-Widom distribution, multimodal distribution.

I. INTRODUCTION

A power system is an electrical energy production and consumption system consisting of power plants, transmission and transformation lines, distribution stations, and loads. Its function is to convert the primary energy found in nature into electrical energy through a power generation device and then supply electrical energy to users through power transmission, transformation, and distribution. From the perspectives of safety and economy, power generation needs to maintain a stable operation status. Even under a very low load demand, power generation still needs to maintain a minimum level of operation. One of the main challenges in power systems is how to properly allocate power generation to deal with diverse and unstable load demands. Based on data reported by the U.S. Energy Information Administration (EIA), domestic customers consumed about 11,000 kilowatt-hours (kWh) on average in 2015. Moreover, electricity demand continues to increase rapidly, putting tremendous pressure on already overloaded power grid infrastructure in the U.S. [1]. However, household electricity consumption differs across different regions and housing types, and the distribution network provides power to a variety of loads, such as residential, industrial, and commercial ones. In particular, the residential load is highly variable compared to other load demands in a power system due to the variety of people's lifestyles [2] and household appliances [3]. Therefore, the capability to handle the impact of load fluctuations on the power grid is an important criterion for evaluating the performance of present and future power systems [1]. In fact, although residential loads are not planned loads, they do change in a cyclic pattern. By studying electricity consumption data, researchers have developed some methods to estimate users' electricity consumption habits, which can help to predict load demands. For example, H. Fell *et al.* developed an empirical approach to estimate residential electricity demand [3]. Lifestyle factors could be one of the key features to predict electricity usage. Income and occupation are also important factors affecting electricity consumption [2]. The ever increasing diversity in the types of generation (new energy sources) and loads

The associate editor coordinating the review of this manuscript and approving it for publication was Emilio Barocio.

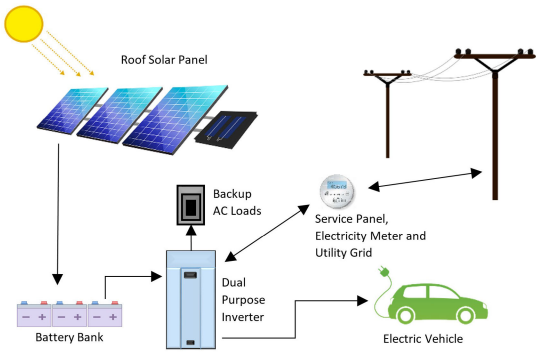


FIGURE 1. Future residential grid. Renewable generations and plug-in devices introduce various uncertainties into power systems.

(plug-in devices, distributed storage) make load demand estimation more challenging than before. In addition, the presence of renewable energy sources and plug-in devices has been transforming the topology of existing power systems significantly as well as changing the power flow direction [4]. As illustrated in Figure 1, the complexity of the new power system topology is increasing due to the many kinds of renewable generations and plug-in electric vehicles, and the fact that the power flow direction is no longer one way. In addition, the uncertainties associate with the resources used to produce renewable energy have an increasing impact on the operation of power systems [5]. For example, in the photovoltaic generation, uncertainty arises from the instability of solar irradiance due to climate changes [3]. In order to preserve the stability and efficiency of a power system, new approaches that can deal with these uncertainties are urgently needed. Most of the engineering problems in a distribution network are subject to uncertainties, especially due to the presence of renewable sources and plug-and-play devices. Uncertainties may also come from daily load variations, generation outages, faults, and failures in power system networks.

Since it is impossible to fully characterize power systems with deterministic methods, stochastic approaches have been attracting more attention. A series of anomaly indicators and fault detection filters have been developed based on the Tracy-Widom distribution [6], [7]. In probabilistic load flow, each input is considered as a random variable in order to incorporate the effects of uncertainties. Common uncertainty factors in conventional generation and new energy generation ignore the uncertainty from load demand, as load demand highly depends on individuals' living patterns [8]. For instance, electric vehicle holding quantity has risen rapidly in recent years, and these vehicles can be considered as storage systems, which will presumably stress the load limits of the network [9]. Consequently, for the effective design and operation of power systems, uncertainty modeling tools are important [31]. Because most loads follow a specific range of variations over time, the load profile of the feeder is a cyclic phenomenon. To better understand the output of these uncertainties, probabilistic approaches are called for. However, most existing load analysis methods are based on

TABLE 1. Approaches of probabilistic load study.

References	Probabilistic model	Contributions
[2], [3]	Unimodal Gaussian	A new point estimation method fused with Rosenblatt transformation to achieve high efficiency.
[31]	Unimodal Gaussian	A new power flow estimation method based on the unscented transformation (UT) which claimed to be accurate and efficient. It can handle correlated Gaussian input random variables.
[6]	Multimodal Gaussian	A Gaussian mixture based robust optimization method for engineering system design claimed high efficiency, accuracy, and convergence.
[11]	non-Gaussian	A novel modeling method for power data based on the random matrix theory (RMT) which can be utilized in a non-Gaussian distribution environment.
Proposed method	Multimodal Tracy-Widom	A new paradigm to model and analyze electricity consumption data based on multimodal Tracy-Widom distribution, which able to characterize correlations in electricity consumption data with multimodal Tracy-Widom approximation.

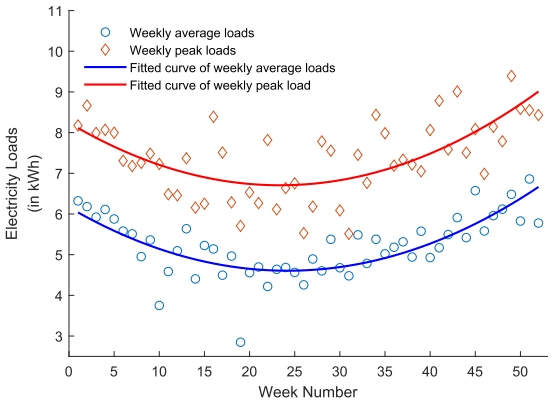


FIGURE 2. Average loads and peak loads of fifty-two weeks. The X-axis is the number of the week, and the Y-axis represents electricity loads. Blue circles are average electricity loads over a week, red diamonds are peak loads in a week. The blue curve and the red curve are fitted curves for weekly average loads and weekly peak loads, respectively.

an assumption that electricity loads are independent random variables and Gaussian distributed [10]. Table 1 summarizes some existing load analysis methods based on different probabilistic model.

Conventional studies of residential electricity loads have focused on the average statistical behavior, which helps a distribution system to allocate the proper amount of electricity to residential loads. On the other hand, the blackout or outage of the grid is often caused by irregular usage behavior or instant demand peak overshoot [11]. As shown in Figure 2, the weekly average loads are concentrated around the blue curve, deterministic approaches are suitable for this case to model and predict the average loads. However, the peak loads (red diamonds) are much more fluctuating than the average loads (blue circles), the conventional methods for

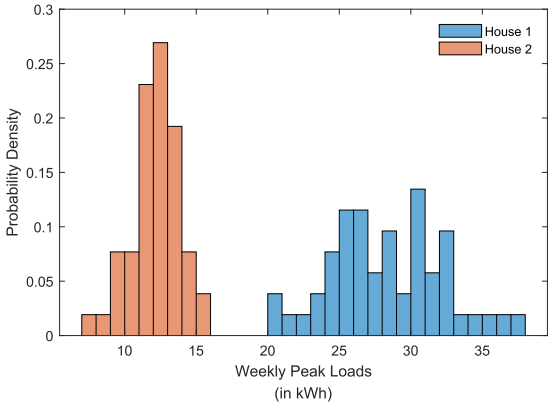


FIGURE 3. Probability density distribution of peak loads of different houses over fifty-two weeks. The X-axis represents the weekly peak loads, and the Y-axis represents the probability density of weekly peak loads.

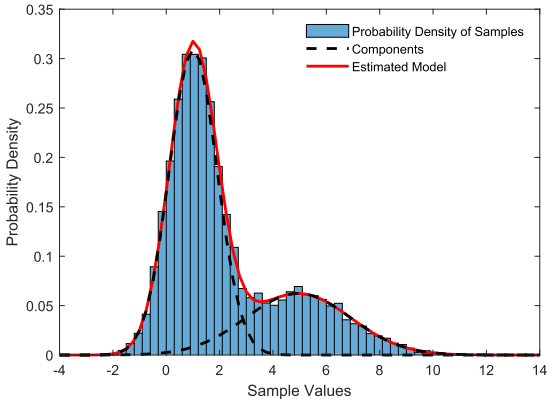


FIGURE 4. Example of multimodal approximation. The X-axis represents the normalized sample values, and the Y-axis represents the probability density. The red solid curve is the estimated multimodal density, the black dashed curves are the Gaussian components of two modes.

average loads study are unable to predict future peak loads. Moreover, the histograms in Figure 3 shows the weekly peak loads of two randomly picked houses, which demonstrates that the variation or differences between houses can be still large. In this figure, house 1's weekly peak loads are much larger than house 2's. Since this distribution is about a data set of an entire year (fifty-two weeks), we can infer that there are some factors that have led consumers to develop different electricity consumption habits. In the Extreme Value Theorem (EVT), the maximum or minimum value will converge to a function if it satisfies certain conditions. Extreme value statistics focuses on the study of the maximum or the minimum of a set of random variables, e.g., residential electricity consumption data or daily water demand. The study of extreme values is very important for time-series problems, and EVT has applications in biology, finance, physics, climate science. However, the extreme value statistics of 'uncorrelated' variables are well studied, strongly correlated random variables have not been well explored [7]. And the residential electricity loads and peak loads are such kind of correlated data which is treated as 'uncorrelated' in previous studies. For example, researchers can use a Gaussian mixture to estimate the probability distribution of a non-Gaussian bus power, the resulting distributions may be unable to reflect asymmetry and tail behavior accurately. However, customers in a residential grid correlated due to some socioeconomic factors.

In real-world scenarios, random variables are more likely to follow a multimodal distribution rather than a unimodal distribution, as is commonly seen in economics, biology, physics, etc. For instance, the absolute magnitude of novae and the size of worker weaver ants are bimodal distributions. Another interesting case is the distribution of daily water demand, which is bimodal, as human activities related to water consumption (e.g. showers, cooking, toilet use) usually peak in the morning and evening periods [12]. Figure 4 shows a density plot of the daily water demand of a settlement, which illustrates the recessive multimodality of the data. Therefore, residential electricity demand can be associated with a bimodal distribution or even a multimodal distribution.

In this article, a new paradigm is proposed to describe extreme values statistics in residential electricity consumption. The major contributions of this article are

- Modeling the maximum residential electricity load as a multimodal distribution of Tracy-Widom distribution components;
- Developing a numerical solver to find the location and scale parameters of a Tracy-Widom distribution.

Unlike the conventional describer (e.g., Gaussian distribution), the proposed Tracy-Widom distribution-based approach reveals that there exists correlation in electricity consumption data. Socioeconomic factors, such as income and occupation play important roles in determining consumption habits, as well as the dependency of random variables. The rest of this article is organized as follows. Section II gives a concise introduction to the Tracy-Widom distribution and multimodal distribution. We then provide a detailed description of the proposed method in Section III. In Section IV, synthesized residential electricity consumption data is used to validate the proposed method. Finally, conclusions and future research plans are outlined in Section V.

II. MATHEMATICAL BACKGROUND

A. MULTIMODAL DISTRIBUTION

A multimodal distribution can be defined as in (1) below, which is commonly expressed in the form of a mixture of possibly different unimodal distributions

$$F(x) = \sum_{i=1}^{n} p_i f_i(x), \tag{1}$$

where $f_i(x)$ is a probability distribution and p_i is the mixing parameter. The mixing parameters satisfies:

$$p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1. \tag{2}$$

The number of modes of the resulting density may not be the same as the number of components. Figure 4 shows an example of using a Gaussian mixture to estimate the distribution of daily electricity loads. The density function of the Gaussian

mixture is stated as in (3) below, where it has five parameters to be determined: μ_1, μ_2 (means), σ_1, σ_2 (variances), and p (mixing parameter).

$$F(x) = p\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + (1-p)\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2} \tag{3}$$

There are some important features worth mentioning when applying a multimodal distribution:

1) A mixture of two Gaussian distributions with the same standard deviation is not bimodal if the difference between the means of the Gaussian components is less than twice the common standard deviation [13].
2) A mixture of two Gaussian distributions with approximately equal mass has a negative kurtosis. The tailness of the mixture is lessened since the two modes on either side of the center of mass offset each other.
3) A mixture of two Gaussian distributions with highly unequal mass has a positive kurtosis. The tailness of the mixture is magnified, as the smaller distribution enhances the tail of the more dominant Gaussian distribution.

Multimodal distributions are very common in both mathematics [14], [15] and natural sciences [13], [16], [17]. Examples of variables with multimodal distributions include the age of a sample of college students and disease patterns. In addition, numerous statistics involving human activities are bimodal distributions, e.g., peak restaurant hours (lunch and dinner), road usage (morning and afternoon rush hours), and residential water/electricity usage (before and after work). According to [18], multimodality can be quantified using a concentration parameter. Most of the previous studies involving multimodal distributions use a Gaussian distribution as the fundamental component. This is because Gaussian distributions play an important role in statistics and can be applied to many different fields. In both natural and social sciences, real-valued random variables with unknown distributions are often assumed as Gaussian distributions. Moreover, the central limit theorem extends Gaussian distributions to another degree. In probability theory, the central limit theorem states that if you have a population and take sufficiently large random samples from the population, then the distribution of the sample means trends toward a Gaussian one, even if the population is not normally distributed.

However, in power systems, energy consumption behavior is affected by a variety of socioeconomic factors [3], which often leads to clustering of the distribution in a population. In addition, the electricity consumption distribution appears asymmetrical due to the limited number of samples [13].

B. TRACY-WIDOM DISTRIBUTION

In 1993, C. Tracy and H. Widom introduced the Tracy-Widom distribution [19], which is the limiting law of the normalized largest eigenvalue of a random Gaussian ensemble. There are three classic random matrix models called the Gaussian ensembles.

- Gaussian Orthogonal Ensemble (GOE, $\beta = 1$): real symmetric matrices;
- Gaussian Unitary Ensemble (GUE, $\beta = 2$): Hermitian matrices;
- Gaussian Symplectic Ensemble (GSE, $\beta = 4$): quaternion matrices.

The Tracy-Widom distribution is an important topic in a wide range of subjects, e.g., mathematical physics [19], [20], economic statistical analysis, ecology, and engineering [7], because it characterizes the inherent randomness of correlated systems, which is universal in real world scenes. A Tracy-Widom distribution is stated in terms of the solution to the Painleve II differential equation. For example, a Tracy-Widom distribution of order 1 is defined as

$$F_2(s) = \exp\left(-\frac{1}{2}\int_s^\infty q(x) + (x-s)\,q^2(x)\,dx\right), \tag{4}$$

where $s \in \mathbb{R}$ and q is the unique solution of the Painleve II differential equation

$$\frac{d^2q(x)}{dx^2}(x) = xq(x) + 2q^3(x) \tag{5}$$

satisfying the boundary condition

$$q(x) \sim Ai(x) \quad as\ x \to +\infty \tag{6}$$

where $Ai(x)$ is the Airy function [19].

The distribution F_2 is associated to a GUE, while Tracy-Widom distributions F_1 and F_4 are associated to a GOE ($\beta = 1$) and a GSE ($\beta = 4$) which are stated in terms of the same Painleve transcendent q:

$$F_1(s) = \exp\left(-\frac{1}{2}\int_s^\infty q(x)dx\right)\sqrt{(F_2(s))} \tag{7}$$

and

$$F_4\left(\frac{s}{\sqrt{2}}\right) = \cosh\left(\frac{1}{2}\int_s^\infty q(x)dx\right)\sqrt{(F_2(s))}. \tag{8}$$

Furthermore, Ramírez *et al.* [21] extended the scope of the Tracy-Widom distributions F_β to all $\beta > 0$.

Theoretically, covariance matrices represent the hidden statistical interdependence structure of the data. In practice, however, sample covariance matrices depend on experimental measurements and cannot reveal the entire interdependence structure [22].

Through Random Matrix Theory (RMT), Johansson [23] and Johnstone [24] introduced a convenient way to study the distribution of the largest eigenvalue of a random matrix. Suppose a random matrix $A \in \mathbb{R}^{m\times n}$ has independent and identically distributed entries. Additionally, the location parameter (μ) and scale parameter (σ) of the corresponding Tracy-Widom distribution are functions of the random matrix's dimensions, and they can be computed using equation (9) and equation (10), respectively.

$$\mu(m,n) = \left(\sqrt{m-1} + \sqrt{n}\right)^2 \tag{9}$$

$$\sigma(m,n) = \left(\sqrt{m-1} + \sqrt{n}\right)\left(\frac{1}{\sqrt{m-1}} + \frac{1}{\sqrt{n}}\right)^{1/3} \tag{10}$$

TABLE 2. Statistics of different types of Tracy-Widom distributions (F_β). The mean (μ_β), variance (σ_β^2), skewness (S_β), and kurtosis (K_β).

β	μ_β	σ_β^2	S_β	K_β
1	-1.206	1.607	0.293	0.165
2	-1.771	0.813	0.224	0.093
4	-2.306	0.517	0.165	0.049

Then, the distribution of the largest eigenvalue λ_1 of a real Wishart matrix AA' in the limit $m, n \to \infty$ with $m/n > 1$, approaches to

$$\frac{\lambda_1 - \mu(m,n)}{\sigma(m,n)} \to W_1 \sim F_1 \tag{11}$$

where F_1 is a Tracy–Widom law of order 1 distribution function.

The numerical evaluations of the Tracy-Widom distributions F_β are very useful in data analysis and other applications. Edelman and Persson [25] were the first to compute numerical solutions to Painleve equations of the types II and V, and their method was able to evaluate eigenvalue distributions of random matrices numerically in MATLAB [25]. Furthermore, these approximation techniques were analytically verified by Bejan [22] and extended as a numerical evaluation method for Painleve II and Tracy-Widom distributions (for $\beta = 1$, 2, and 4) in S-PLUS [22]. After that, a fast algorithm was proposed by Bornemann [26] to numerically evaluate the Tracy-Widom distribution and the density functions $f_\beta(s) = dF_\beta/ds$ for all three Gaussian ensembles ($\beta = 1, 2, 4$). These algorithms are useful for calculating statistics, such as the mean, variance, skewness, and excess kurtosis, of F_β numerically. The results are shown in Table 2. Although these numerical statistics provide an overall profile of the density functions of a Tracy-Widom distribution, Chiani (2014) has also provided a simpler and faster way to approximate F_β based on a shifted gamma distribution [27].

III. MIXTURE OF THE TRACY-WIDOM DISTRIBUTION

A. PERFORMANCE METRICS

A statistical distance is a quantified measurement that describes the distance or difference between two statistical objects (e.g., random variables, probability distributions, samples). There are many different statistical distances in statistics, probability theory, and information theory; however, not all statistical distances are metrics, as some of them lack one or more axioms belonging to proper metrics. A statistical distance can be considered a metric only if it satisfies the following conditions:

1) non-negativity: $D(f_1, f_2) \geq 0$
2) identity of indiscernibles: $D(f_1, f_2) = 0$ if and only if $x = y$
3) symmetry: $D(f_1, f_2) = D(f_2, f_1)$
4) subadditivity: $D(f_1, f_3) \leq D(f_1, f_2) + D(f_2, f_3)$,

where D is a statistical distance function, and f_i, $i = 1, 2, 3$ are probability distributions. That said, some commonly used statistical distances, e.g. f-divergence and energy distance, are not suitable to measure the performance difference between two probability densities in this study. Since we are measuring the performance of a density approximation method, the disparity between the original density and estimated density is a proper performance metric. The disparity between two densities d is defined as

$$d = 1 - \int_{-\infty}^{+\infty} min(f(x), \hat{f}(x))dx. \tag{12}$$

where $f(x)$ is the original PDF, and $\hat{f}(x)$ is the estimated PDF. In this way, the density approximation can be transformed as an optimization problem to minimize d.

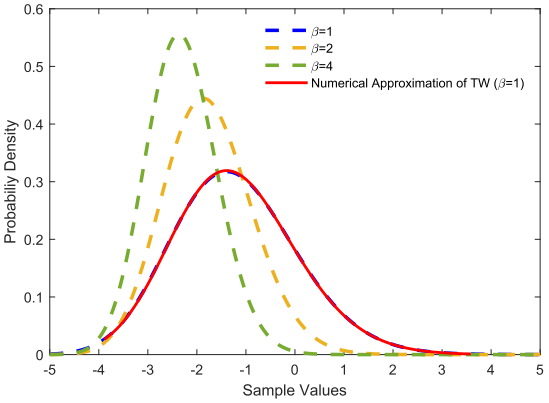


FIGURE 5. Tracy-widom distribution and numerical approximation. $\beta = 1, 2$ and 3 associated with density functions (4),(7), and (8), respectively. The red solid line is a numerical approximation of type I.

B. ESTIMATING PARAMETERS OF THE TRACY-WIDOM DISTRIBUTION

Although we mentioned in III-A that some optimization methods are available to approximate probability density, the analytical form of a Tracy-Widom distribution is quite complicated. Solving equations (4-6) inside of an optimization problem is not efficient and is not even applicable in practice. Moreover, existing numerical methods [23], [24] are not applicable to our case, though it is worth noting that our investigation, we found A. Bejan's [22] numerical approximation method of the Tracy-Widom distribution can achieve high accuracy. Figure 5 shows the numerical approximation of type I Tracy-Widom distribution, the red solid line coincides with the blue dash line, which means this numerical approximation is a reliable benchmark for our application.

Fortunately, there exist in the literature some numerical evaluations of Tracy-Widom distributions [22], [25], [26] that one can use. Inspired by the works of Johansson [23] and Jacob *et al.* [28], by finding a proper pair of location parameter μ and scale parameter σ, we can apply a certain type of Tracy-Widom distribution, e.g., F_1, to arbitrary data, not just eigenvalues of a random matrix. Here, we have formulated a quadratic programming (QP) problem to solve for μ and σ of type-I Tracy-Widom distribution F_1. The QP problem is defined as follows:

$$\min_{\mu,\sigma} \; d = 1 - \int_{-\infty}^{+\infty} min\left(\hat{f}_{TW}(x, \mu, \sigma), f(x)\right) dx$$
$$\text{s.t.} \int_{-\infty}^{+\infty} \hat{f}_{TW}(x, \mu, \sigma)dx = 1. \tag{13}$$

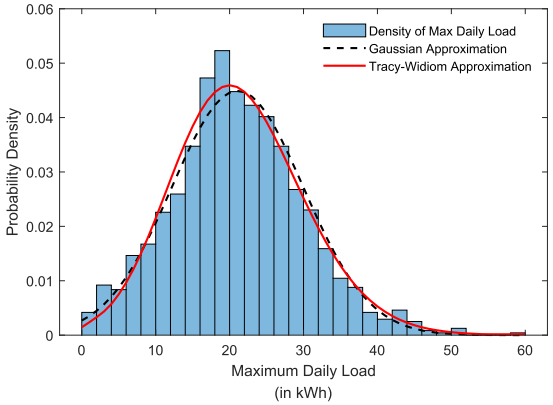


FIGURE 6. **Unimodal distribution approximation. This histogram is the probability density of the maximum daily load of a settlement. The red solid line and the black dash line represent the estimated densities of maximum daily load by Tracy-Widom distribution and Gaussian distribution, respectively.**

where $\hat{f}_{TW}(x)$ is the estimated density of the type-I Tracy-Widom distribution, and $f(x)$ is the original density of the data set. We use the numerical result from Bejan [22] as a benchmark; then the QP problem will search for a pair of μ and σ, where $\hat{f}_{TW}(x)$ fits the shape of the original density of the data. Figure 6 shows the result of Tracy-Widom density approximation to an electricity load data set and the approximation performances are shown in Table 2. From Figure 6 we can see that Tracy-Widom approximation is better than Gaussian approximation at preserving the tailedness and asymmetry of the original density with a very high approximation accuracy.

This Tracy-Widom approximation method is fast and easy to apply. The result is also promising in general, since the randomness of power system data is not independent in practice. That is, behind the data, there are a lot of human interactions that could be highly correlated due to various undiscovered factors. Thus, a Tracy-Widom distribution is much better at characterizing correlated data than a Gaussian distribution. Therefore, the algorithm searches for the appropriate parameters that maximize the overlapping area between the original data density and the approximated Tracy-Widom density does not a require specific data format. The algorithm is described in greater detail below.

C. TRACY-WIDOM MIXTURE

As mentioned in section II, the statistics of power system data are often bimodal or multimodal. In addition, it is impossible to apply unimodal distribution to a multimodal case. Traditional multimodal distributions employ a Gaussian distribution as their component. However, the density of power data often appears highly skewed and heavy-tailed, especially given extreme values like peak load demands. Thus naturally we fused a Tracy-Widom distribution into a multimodal distribution, as a Tracy-Widom distribution is more suitable in describing correlated data.

Instead of using a Gaussian distribution as the component in the multimodal distribution, we used a Tracy-Widom distribution, such that $f_i(x) = F_\beta$ in equation (1). Then, multimodal

Algorithm 1: Approximation to Tracy-Widom Distribution

Result: Scale (σ) and location (μ) Parameters of Tracy-Widom distribution
compute and plot the reference PDF and CDF of TW,
make initial guess of σ_0 and μ_0;
while *d in equation (13) is greater than threshold ϵ* **do**
  step 1: making a guess of σ_i and μ_i, and plug them into equation(11);
  step 2: using a conventional interpolation method to obtain a smooth and continuous representation of Tracy-Widom distribution $\hat{f}_i(x, \mu_i, \sigma_i)$;
  step 3: plugging the data x into $\hat{f}_i(x, \mu_i, \sigma_i)$, compute the density approximation $\hat{f}_i(x, \mu_i, \sigma_i)$;
  step 4: computing disparity d_i between $\hat{f}_i(x, \mu_i, \sigma_i)$ and density of data $f(x)$;
  if $d_i > \epsilon$ **then**
    Update (make a new guess) σ_i and μ_i;
    go to step 1;
  else
    terminate algorithm and output $\sigma = \sigma_i$ and $\mu = \mu_i$;
  end
end

distribution (1) can be rewritten as equation (14):

$$F_{TW}(x) = pf_{TW1}(x) + (1-p)f_{TW2}(x), \tag{14}$$

where $F_{TW}(x)$ is a bimodal Tracy-Widom distribution, $f_{TW1}(x), f_{TW2}(x)$ are Tracy-Widom distributions with different location and scale parameters, and p is the mixing parameter.

Therefore, we formulated a QP problem to approximate the Tracy-Widom mixture (14) as follows:

$$\begin{aligned} &\min_{\mu_i,\sigma_i,p} \; 1 - \int_{-\infty}^{+\infty} min\left(\hat{F}_{TW}(x, \mu_i, \sigma_i, p), F(x)\right) dx \\ &\text{s.t.} \int_{-\infty}^{+\infty} \hat{f}_{TWi}(x)dx = 1, \quad \sigma_i \geq 0, \forall i, \end{aligned} \tag{15}$$

where $\hat{F}_{TW}(x, \mu_i, \sigma_i, p)$ is the estimation density of the Tracy-Widom mixture, $F(x)$ is the original density of the data, $\hat{f}_{TWi}(x)$, $i = 1, 2$ are unimodal Tracy-Widom distributions, σ_i are scale parameters of the Tracy-Widom components. Similar to the former QP problem of the Tracy-Widom approximation in III-B, the multimodal Tracy-Widom approximation will find the scale and location parameters and the mixing parameters numerically. This is a data driven method, applicable to any other data sets beyond power data.

IV. CASE STUDY

In this section, electricity consumption data of a settlement that contains 1000 households is used as a case study to demonstrate the performance of the proposed algorithm. The electricity profile was generated using the load profile generator (LPG) proposed in [29], which is available

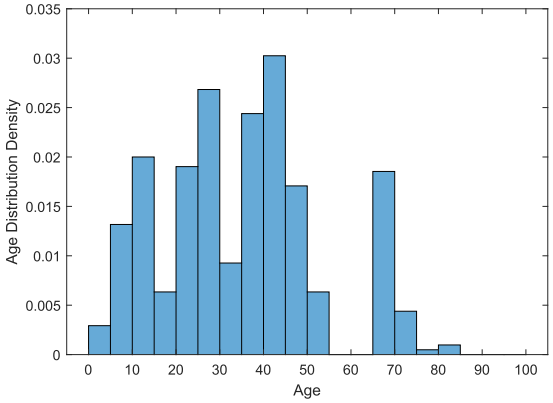


FIGURE 7. Population age distribution. A young settlement, more than 90% of residents are less than 60.

TABLE 3. Accuracy of unimodal distribution approximation.

Methods	Tracy-Widom approximation	Gaussian approximation
Accuracy	99.80%	99.25%

TABLE 4. Income status. The number of employees of a house will affect the regularity of work and rest and consumption habits.

Family Sizes	Number of Earner	Percent
1	0	15.10%
1	1	25.70%
2	0	12.70%
2	1	21.70%
3	0	4.60%
3	1	4.00%
3	2	3.80%
4	0	3.30%
4	1	2.90%
4	2	2.80%
≥5	0	1.30%
≥5	1	1.10%
≥5	2	1.00%

in [30]. The LPG is a modeling tool for residential energy consumptions of individual households and large settlements. The profile generator simulates a full behavior of people in a household and is based on the desired model, which includes typical operation patterns for more than 100 electrical devices. All traits of a household model, including house size, family size, ages, occupations, living pattern, and appliance activities, can be customized by the user. The dataset investigated in this article was generated based on the climate and geographic profile in Germany. The household model factors, such as age, family size, house size, occupations, and living pattern, are based on the average distribution of the German population. The age and income statistics of the dataset are shown in Figure 7 and Table 4, respectively. Obviously, this is a young settlement, most of them are young or middle-aged couples with children. Also, more than 50% of families have more than one earner. The climate profile and location of all households are set identically in order to present consistent results.

Electricity consumption patterns widely differ among customers and are highly stochastic. Figure 8 shows two

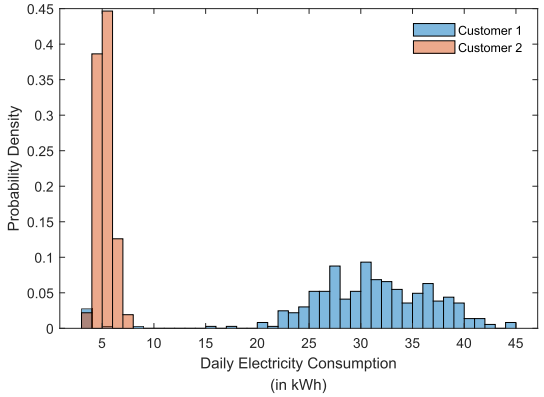


FIGURE 8. Example of household daily electricity loads. The X-axis represents the daily electricity consumption of customers, the Y-axis represents the probability density of customers' consumption data.

individual customers' electricity load profiles, over a year. Customer 1's data is concentrated in a small range and it shows customer 1 has an energy-saving lifestyle. On the contrary, customer 2's daily electricity consumption behavior varies substantially and is hard to predict, as the probabilities are evenly distributed on a wide sample range. In addition to characterizing a data set, skewness and kurtosis give deeper insight into the behavior. In statistical analyses, skewness is a measurement that quantifies the asymmetry of a probability distribution about its mean. The skewness of a real-valued random variable X is defined as the third standardized moment $\tilde{\mu}_3$:

$$\tilde{\mu}_3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}\left[(X-\mu)^2\right])^{3/2}}, \quad (16)$$

where μ is the mean, σ is the standard deviation, E is the expectation operator, and μ_3 is the third central moment. Similar to skewness, kurtosis describes the shape of a probability distribution and its tail behavior. The kurtosis is the fourth standardized moment is defined as

$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}, \quad (17)$$

where μ_4 is the fourth central moment.

The skewness and kurtosis of the data distribution, the Tracy-Widom distribution and the Gaussian distribution are listed in Table 5. From the skewness column, it is obvious that the Tracy-Widom distribution is more appropriate for depicting the asymmetry of the data distribution than the Gaussian distribution. The kurtosis shows that the data distribution is heavy-tailed a trend the Tracy-Widom distribution manages to capture. In summary, the residential electricity distribution and the Tracy-Widom distribution are asymmetric and right-tailed. In contrast, the Gaussian distribution is symmetric, making the Tracy-Widom distribution better than the Gaussian distribution. Furthermore, as shown in Figure 9, the density of the max daily load (blue histogram) has two modes and is slightly skewed to the right in each mode. Thus, the result corresponding to the hypothesis is that a finite data

TABLE 5. Skewness and excess kurtosis of the electricity consumption data, Tracy-Widom distribution, and Gaussian distribution.

Density Type	Skewness	Excess kurtosis
Data	0.3114	3.3601
Tracy-Widom Distribution	0.2934	0.1652
Gaussian Distribution	0	0

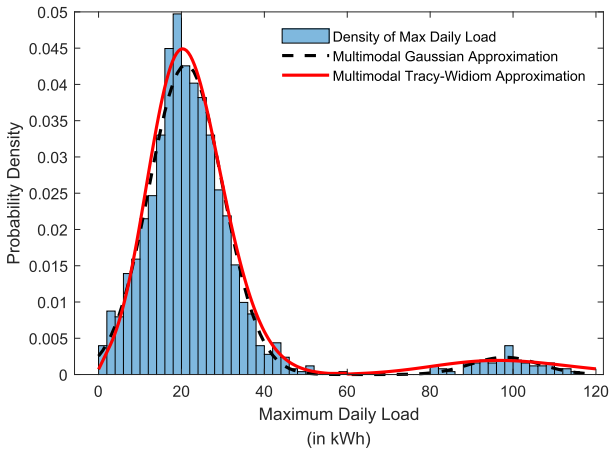


FIGURE 9. Multimodal distribution approximation. The red solid line and the black dash line represents the estimated densities of maximum daily load by Tracy-Widom mixture and Gaussian mixture, respectively.

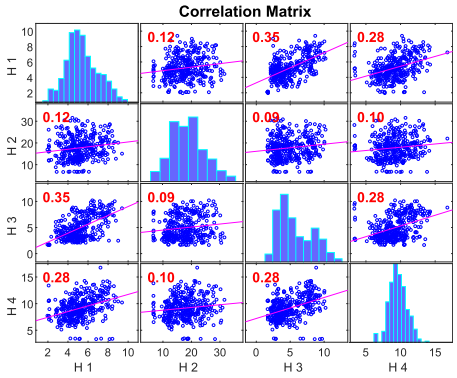


FIGURE 10. Correlation matrix of daily loads. H 1-4 represent four random houses in a settlement. The histograms in the main diagonal are the probability densities of their daily electricity consumption. The plots in off-diagonal show the correlation between each pair of houses. The correlation coefficients highlighted in red indicate which pairs of houses have correlation coefficients significantly different from zero.

sample trends toward a Tracy-Widom distribution rather than a Gaussian distribution.

A. ANALYSIS

Most existing research on load demands are concentrated on average values; however, extreme value analysis (e.g., peak load demands) is critical to power system operational safety, since an unexpected load demand impulse may a cause grid blackout. Classical EVT is committed to statistical analysis of the maximum (or minimum) of a set of uncorrelated random variables [28]. Furthermore, the conventional study of load demand always considers a customer as an independent individual, when in fact, in most real-world systems, e.g., power systems, the underlying random variables are typically correlated. In Figure 9, the multimodal Tracy-Widom

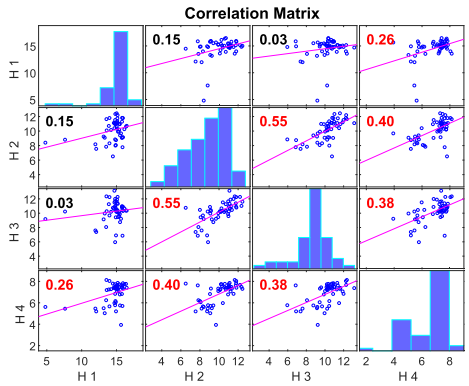


FIGURE 11. Correlation matrix of weekly loads. H 1-4 represent four random houses in a settlement. The correlation coefficients highlighted in red indicate which pairs of houses have correlation coefficients significantly different from zero.

TABLE 6. Accuracy of density estimation by Tracy-Widom mixture and Gaussian mixture.

Methods	Tracy-Widom mixture	Gaussian mixture
Accuracy	99.27%	93.31%

TABLE 7. Accuracy of major mode approximation by multimodal Tracy-Widom distribution and GEVD.

Methods	Tracy-Widom distribution	GEVD
Accuracy	99.80%	99.73%

approximation is more accurate than multimodal Gaussian approximation. On the left part of the data density, it is asymmetric and skewed to its left, the multimodal Tracy-Widom approximation captures the skewness. In terms of overall performance, also demonstrated in Table 4, Tracy-Widom mixture has 99.27% accuracy in estimating the density of this data set. Compare to 93.31% accuracy by the Gaussian mixture approximation, there are some undetermined factors affecting electricity consumption behavior, which could be socioeconomic factors like income, education, etc. Even though human activities are random, the daily life patterns of individuals are clustered in a population by age, educational background, etc. Figure 10 and 11 are daily and weekly electricity correlation maps of four random houses respectively. The correlation coefficients highlighted in red, e.g., 0.28, 0.55, show there are correlations between power consumption patterns of houses. Additionally, the correlations in the electricity consumption data indicate that the independent random variable assumption inherited in conventional electricity load research works [5], [8], [9] is inappropriate for the modern electricity consumption model. Currently, we do not have enough dataset that includes people's socioeconomic information to allow us to explore the impact of these factors on electricity consumption. What one can tell that the Tracy-Widom mixture is better than the Gaussian mixture as seen from Figure 9. This implies that residential electricity consumption data is correlated, as are the maximum demands.

Moreover, we intercepted the left part or major mode of the data density, and applied proposed Tracy-Widom density approximation and the generalized EVD to estimate

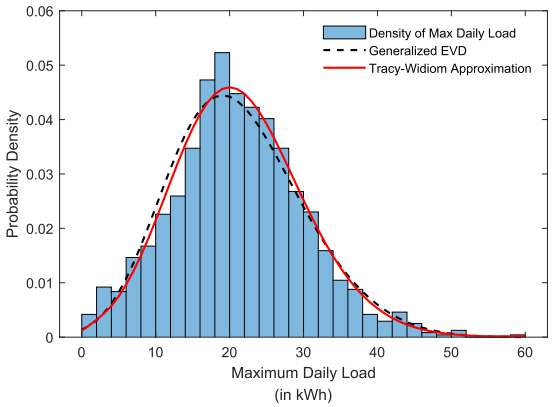


FIGURE 12. Compare Tracy-Widom distribution with generalized extreme value distribution (GEVD). The red solid line is a Tracy-Widom density approximation, the black dash line is a density approximation by generalized extreme value distribution.

the density of the major mode. As shown in Table 6, both approximations achieved very promising results: 99.80% and 99.73% accuracy for Tracy-Widom approximation and generalized EVD approximation respectively. And from Figure 11, the only difference is regarding the skewness: generalized EVD approximation is over skewed to the left, but Tracy-Widom approximation is more consistent with the data density. The reason behind this result is that EVD is used to describe the Poisson point process, of which samples are independent. Quite the contrary, the Tracy-Widom distribution is used to describe a strongly coupled point process, of which samples are dependent or correlated. Therefore, given that there are some correlations between variables, the traditional extreme value distribution may not be appropriate in describing the statistics of this residential electricity data. As such, a Tracy-Widom distribution that represents the underlying correlation factors well, is used here as a better solution. Figure 12 shows the approximations of the major mode using the Tracy-Widom distribution and extreme value distribution. As can be seen, the Tracy-Widom distribution is more proper in representing the peak location and skewness of data without much loss of accuracy. From this perspective, the proposed Tracy-Widom mixture approach is more appropriate than the Gaussian mixture to estimate the probability density of residential electricity load and peak load. In addition, this approach can be easily applied to other data set when correlation exists.

V. CONCLUSION AND FUTURE WORK

In this article, we investigated residential electricity consumption data to predict the potential excess load summit as it is crucial for a network to allocate appropriate loads in advance to prevent a large area blackout. We demonstrated the new framework's performance considering a residential settlement with 1000 households. We found that certain socioeconomic factors, such as living patterns, have a significant impact on electricity usage behavior. Furthermore, the multimodality of the maximum electricity consumption is also determined by some socioeconomic factors. Our results show that a Tracy-Widom distribution is more reliable than a Gaussian distribution in modeling maximum electricity consumption data in a multimodal fashion. A potential limitation of this study is the proposed method is only compared with the most commonly used methods, like the Gaussian approximation and GEVD approximation, but there are a lot other distribution functions potentially better than Tracy-Widom distribution. We cannot cover all the comparison in this article due to the limitation of time and energy.

The proposed work is tested and validated with a simulated load profile, and it has a promising performance on the simulated data. Currently, we do not have real load data available that associate with customer's socioeconomic factors, e.g., income, age, daily working hours. In the future, real load data will be used to evaluate the proposed method to further justify its performance. Additionally, we also like to identify the list of socioeconomic factors that affects electricity usage behavior in our future research. Most statistical studies on residential loads are based on the assumption that households (entries) in the residential network are statistically independent and identically distributed. However, our results show that entries in a residential network are correlated and clustered. These underlying factors are essential and crucial to power supply decisions. Moreover, quantitative metrics could arise to describe and measure these factors, which would precise tools for cost reductions for both the generation side and customers. Furthermore, some large scale anomaly events can change the weights of socioeconomic factors associated with electricity consumption. For example, during the recent pandemic, the residential electricity load is increased significantly since people find a new way to live and work to adjust to the national wide lockdown. Identifying the impact of major anomaly events on electricity consumption and providing a new power supply strategy to mitigate economic loss will also be our future tasks.

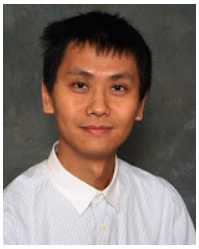
From the results shown in section IV, the proposed method has a better performance in modeling the daily max electricity consumption per house than conventional methods. With a more precise model, the peak load-interval can be further narrow down. This will help a network to allocate appropriate loads in advance to prevent a large area blackout. On the other hand, by discovering the correlations in electricity consumption data, underlying factors, e.g., income, will be further characterized in future research. Eventually, a comprehensive model of electricity consumption behavior will benefit both customers and suppliers.

REFERENCES

[1] W. Su, J. Wang, and D. Ton, "Smart grid impact on operation and planning of electric energy systems," in *Handbook Clean Energy Systerm*, J. Yan, Ed. Hoboken, NJ, USA: Wiley, 2015.
[2] T. F. Sanquist, H. Orr, B. Shui, and A. C. Bittner, "Lifestyle factors in U.S. residential electricity consumption," *Energy Policy*, vol. 42, pp. 345–364, Oct. 2012.
[3] H. Fell, S. Li, and A. Paul, "A new look at residential electricity demand using household expenditure data," *Int. J. Ind. Org.*, vol. 33, pp. 37–47, Dec. 2014.
[4] H. Pourbabak, T. Chen, B. Zhang, and W. Su, "Control and energy management system in microgrids," in *Proc. Clean Energy Microgrids*, 2017, pp. 109–133.

[5] X. Xu and Z. Yan, "Probabilistic load flow calculation with quasi-Monte Carlo and multiple linear regression," *Elect. Power Energy Syst.*, vol. 88, pp. 1–12, Oct. 2017.
[6] Z. L. Huang, J. W. Zhang, T. Kumar, T. G. Yang, S. G. Deng, and F. Y. Li, "Robust optimization for micromachine design problems involving multimodal distributions," *IEEE Access*, vol. 7, pp. 91838–91849, 2019.
[7] Y. Yan, G. Sheng, R. C. Qiu, and X. Jiang, "Big data modeling and analysis for power transmission equipment: A novel random matrix theoretical approach," *IEEE Access*, vol. 6, pp. 7148–7156, 2018.
[8] B. R. Prusty and D. Jena, "A critical review on probabilistic load flow studies in uncertainty constrained power systems with photovoltaic generation and a new approach," *Renew. Sustain. Energy Rev.*, vol. 69, pp. 1286–1302, May 2017.
[9] P. Amid and C. Crawford, "A Cumulant-Tensor-Based probabilistic load flow method," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5648–5656, Sep. 2018.
[10] H. Alfredo and W. H. Tang, "Probability concepts in engineering," *Planning*, vol. 1, pp. 1–5, Jul. 2007.
[11] B. Han, L. Luo, G. Sheng, G. Li, and X. Jiang, "Framework of random matrix theory for power system data mining in a non-Gaussian environment," *IEEE Access*, vol. 4, pp. 9969–9977, 2016.
[12] S. Buchberger, T. Omaghomi, T. Wolfe, J. Hewitt, and D. Cole, "Peak water demand study," Univ. Cincinnati, Cincinnati, OH, USA, Tech. Rep. 33087, 2017.
[13] M. F. Schilling, A. E. Watkins, and W. Watkins, "Is human height bimodal?" *The Amer. Stat.*, vol. 56, pp. 223–229, Dec. 2002.
[14] C. V. Fiorio, V. A. Hajivassiliou, and P. C. B. Phillips, "Bimodal-ratios: The impact of thick tails on inference," *Econometrics J.*, vol. 13, pp. 271–289, Aug. 2010.
[15] A. M. Carolan and J. C. W. Rayner, "One sample tests for the location of modes of nonnormal data," *J. Appl. Math. Decis. Sci.*, vol. 5, no. 1, pp. 1–19, Jan. 2001.
[16] R. Sanjuan, "Mutational fitness effects in RNA and single-stranded DNA viruses: Common patterns revealed by site-directed mutagenesis studies," *Phil. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 365, pp. 1975–1982, 2010.
[17] A. Eyre-Walker and P. D. Keightley, "The distribution of fitness effects of new mutations," *Nature Rev. Genet.*, vol. 8, pp. 610–618, Sep. 2007.
[18] P. C. B. Phillips, "A remark on bimodality and weak instrumentation in structural equation estimation," *Econ. Theory*, vol. 22, p. 5, Oct. 2006.
[19] C. A. Tracy and H. Widom, "Level-spacing distributions and the Airy kernel," *Commun. Math. Phys.*, vol. 159, no. 1, pp. 151–174, 1994.
[20] C. A. Tracy and H. Widom, "The distributions of random matrix theory and their applications," in *New Trends Math. Phys.*, V. Sidoravicius, Ed., Dordrecht, Netherlands: Springer, vol. 2009, pp. 753–765.
[21] J. A. Ramãrez, B. Rider, and B. Virãig, "Beta ensembles, stochastic Airy spectrum, and a diffusion," *J. Amer. Math. Soc.*, vol. 24, no. 4, pp. 919–944, 2011.
[22] A. I. Bejan, "Largest eigenvalues and sample covariance matrices," Univ. Warwick, Coventry, U.K., Tech. Rep., 2005.
[23] K. Johansson, "Shape fluctuations and random matrices," *Commun. Math. Phys.*, vol. 209, pp. 437–476, Oct. 2000.
[24] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, vol. 29, pp. 295–327, Dec. 2001.
[25] A. Edelman and P. Persson, "Numerical methods for eigenvalue distributions of random matrices," Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 0501068, 2005.
[26] F. Bornemann, "On the numerical evaluation of distributions in random matrix theory: A review," 2009, *arXiv:0904.1581*. [Online]. Available: http://arxiv.org/abs/0904.1581
[27] M. Chiani, "Distribution of the largest eigenvalue for real Wishart and Gaussian random matrices and a simple approximation for the Tracy-Widom distribution," *J. Multivariate Anal.*, vol. 129, pp. 69–81, Oct. 2014.
[28] M. Jacob, C. Neves, and D. V. Greetham, "Extreme value statistics," in *Forecasting Assessing Risk Individual Electrics Peaks*, Cham, Switzerland: Springer, 2019, pp. 61–84.
[29] N. Pflugradt, "Modellierung von Wasser und Energieverbräuchen in Haushalten," Ph.D. dissertation, Dept. Mech. Eng., Technischen Univ. Chemnitz, Chemnitz, Germany, 2016.
[30] N. Pflugradt. (Jul. 2020). *LoadProfileGenerator*. [Online]. Available: https://www.loadprofilegenerator.de/
[31] M. Aien, M. Fotuhi-Firuzabad, and F. Aminifar, "Probabilistic load flow in correlated uncertain environment using unscented transformation," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 2233–2241, Nov. 2012.

SHIWEN LIAO (Student Member, IEEE) received the B.S. degree in applied physics from the Huazhong University of Science and Technology, Wuhan, China, in 2011, and the M.Sc. degree in electrical engineering from the University of Michigan-Dearborn, Dearborn, USA, in 2014, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests include data mining of electricity load data, consensus-based distributed control of multi-agent systems, and big data analytics for fault diagnosis.



LU WEI (Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University, China, in 2006, the M.Sc. degree (Hons.) from the Helsinki University of Technology, Finland, in 2008, and the D.Sc. degree (Hons.) from Aalto University, Finland, in 2013. He held a Postdoctoral position at the Department of Mathematics and Statistics, University of Helsinki, Finland, from 2013 to 2015, and the School of Engineering and Applied Sciences, Harvard University, USA, from 2015 to 2016. Since Fall 2016, he has been an Assistant Professor with the Department of Electrical and Computer Engineering, University of Michigan-Dearborn, USA. His research interest includes random matrix theory with applications to communications theory and information theory.

TAEHYUNG KIM (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2003. He was a Senior Research Engineer with the Digital Appliances Research Center, Samsung Electronics, in 2003. From 2004 to 2005, he was a Postdoctoral Researcher with the Advanced Vehicle, Power Electronics, and Motor Drive Laboratory, Texas A&M University. In 2005, he joined the Department of Electrical and Computer Engineering, University of Michigan-Dearborn, where he is currently an Associate Professor. His research interests include electric and hybrid electric vehicles, power electronic, and motor drives. He was a recipient of the 2012 Second Prize Paper Award from the IEEE Industry Applications Society (Annual Society's Best Magazine Article).

WENCONG SU (Senior Member, IEEE) received the B.S. degree (Hons.) from Clarkson University, Potsdam, NY, USA, in 2008, the M.S. degree from Virginia Tech, Blacksburg, VA, USA, in 2009, and the Ph.D. degree from North Carolina State University, Raleigh, NC, USA, in 2013. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Michigan-Dearborn, USA. He is also a Registered Professional Engineer (P.E.) at MI, USA. His current research interests include power systems, electrified transportation systems, and cyber-physical systems. He is a Fellow of IET. He was a recipient of the 2015 IEEE Power and Energy Society (PES) Technical Committee Prize Paper Award and the 2013 IEEE Industrial Electronics Society (IES) Student Best Paper Award. He is an Editor of the IEEE TRANSACTIONS ON SMART GRID and an Associate Editor of IEEE ACCESS and the IEEE DataPort.