Extreme Photovoltaic Power Analytics for Electric Utilities

Zefan Tang, *Student Member, IEEE*, Peng Zhang, *Senior Member, IEEE*, Kunihiro Muto, Martial Sawasawa, Marissa Simonelli, *Student Member, IEEE*, Christopher Gutierrez, *Student Member, IEEE*, Jaemo Yang, Marina Astitha, David A. Ferrante, Joseph N. Debs, Robert Manning, *Member, IEEE*, and James Mader

***Abstract*—Obtaining high-fidelity information on extreme photovoltaic (PV) power is critical for electric utility system planning and operations. However, a scarcity of extreme data has previously made achieving an accurate estimate of extreme PV power an intractable challenge. In response to this challenge, this paper presents Extreme PV Power Analytics (EPVA). It utilizes k-means clustering to determine which PV systems have similar behaviors in their extreme capacity factors (ECFs) in order to incorporate more extreme data in an extreme value analysis. This extreme value analysis is subsequently applied to obtain the distribution of ECFs. Zone partitioning results and ECF distribution results for The United Illuminating Company service territory are presented to validate the effectiveness and efficacy of EPVA.**

***Index Terms*—Extreme photovoltaic power analytics, electric utility, k-means clustering, extreme value analysis.**

NOMENCLATURE

$\bar{r}_s, \bar{r}_h, \bar{r}_t, \bar{r}_w, \bar{r}_c$	The average correlation coefficient of each meteorological variable.
$\bar{S}$	The average Silhouette value.
$\bar{X}, \bar{Y}$	The average of X_i / Y_i in the N-data set.
μ_i	The centroid of cluster C_i.
$\mu_{is}, \mu_{ih}, \mu_{it}, \mu_{iw}, \mu_{ic}$	Each meteorological variable of the centroid of cluster C_i.
ξ_u	The probability of exceeding u.
b_j	The lowest average distance between weather site x_j and the other weather sites in any other cluster.
C_i	The ith cluster.
C_{ft}	PV capacity factor at time t.
d_j	The average distance between weather site x_j and other weather sites in the same cluster.
$E_{cf\Delta t}$	Extreme capacity factor at a given time interval from t_1 to t_n.
k	The number of clusters in k-means clustering.
m_s	The number of total samples.
m_u	The number of the samples exceeding u.
N	The number of total data points within a given time interval for a single PV site.
N_1	The number of PV sites available for calculating correlation coefficients.
$n_1, n_2, \ldots, n_k$	The number of weather sites in clusters $C_1, C_2, \ldots, C_k$.
N_2	The number of groups of meteorological variables within a given time interval for weather site x_j.
N_3	The number of total weather sites.
n_y	The number of samples each year.
P_t, P_r	Output AC power at time t / AC nameplate rating.
r	Correlation coefficient between PV output power and one meteorological variable.
S_j	The Silhouette value of weather site x_j.
u, α, γ	Location / Scale / Shape.
V	The within-cluster sum of squares (WCSS).
W_s, W_h, W_t, W_w, W_c	Each meteorological weight.
X_i, Y_i	PV output power / One meteorological variable of the ith data point.
x_j	The jth weather site within cluster C_i.
x_r, z_N	Return level / N-year return level.
$x_{jsni}, x_{jhni}, x_{jtni}, x_{jwni}, x_{jcni}$	The ith normalized meteorological variables of weather site x_j.

Manuscript received June 18, 2018; revised October 14, 2018; accepted November 24, 2018. Date of publication December 3, 2018; date of current version December 18, 2019. This work was supported in part by the Eversource Energy Center under Grants 6200980 and 6200990, and in part by the National Science Foundation under Award Nos. CNS-1647209 and ECCS-1831811. Paper no. TSTE-00581-2018. *(Corresponding author: Peng Zhang.)*

Z. Tang, P. Zhang, K. Muto, M. Sawasawa, M. Simonelli, and C. Gutierrez are with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269-4157 USA (e-mail: zefan.tang@uconn.edu; peng.zhang@uconn.edu; kunihiro.muto@uconn.edu; martial.sawasawa@uconn.edu; marissa.simonelli@uconn.edu; christopher.gutierrez@uconn.edu).

J. Yang and M. Astitha are with the Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT 06269-3037 USA (e-mail: jaemo.yang@uconn.edu; marina.astitha@uconn.edu).

D. A. Ferrante and J. N. Debs are with Eversource Energy, Berlin, CT 06037 USA (e-mail: david.ferrante@eversource.com; joseph.debs@eversource.com).

R. Manning and J. Mader are with The United Illuminating Company, Orange, CT 06477 USA (e-mail: Robert.Manning@uinet.com; Jim.Mader@uinet.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSTE.2018.2884500

$x_{jsn}, x_{jhn}, x_{jtn}, x_{jwn}, x_{jcn}$	Normalized values of x_{js}, x_{jh}, x_{jt}, x_{jw}, x_{jc}.
$X_{js}, X_{jh}, X_{jt}, X_{jw}, X_{jc}$	The representative meteorological variables for weather site x_j in the calculation of Euclidean distance in k-means clustering.
$x_{js}, x_{jh}, x_{jt}, x_{jw}, x_{jc}$	Each meteorological variable of weather site x_j.
$x_{sa}, x_{ha}, x_{ta}, x_{wa}, x_{ca}$	The maximum value of each meteorological variable among the data from all weather sites within a given time interval.
$x_{si}, x_{hi}, x_{ti}, x_{wi}, x_{ci}$	The minimum value of each meteorological variable among the data from all weather sites within a given time interval.

I. Introduction

POWER distribution grids in the United States are being impacted by the increasingly deep integration of photovoltaic (PV) plants [1]–[3]. For instance, The United Illuminating Company (UI), which distributes power to 17 towns and cities in Connecticut, has approximately 7,000 residential and 300 commercial or industrial PV generators interconnected to its system with average nameplate ratings of 6.27 kW and 108 kW, respectively. However, high PV output can have a serious impact on grid systems and customers [4], [5]. One common challenge faced by most electric utilities is the overvoltage caused by extremely high PV output power. According to the American National Standards Institute (ANSI), the service voltage must not exceed the upper limit of 5% above the nominal for longer than one minute. The Public Utility Regulatory Authority (PURA), which regulates Connecticut utilities, mandates an even stricter service voltage upper limit of 3% above nominal [6].

UI conducts a screening process for all PV interconnection applications to ensure that all regulatory requirements are satisfied and all safety concerns are addressed. System models are created to check for potential overvoltage based on the magnitude of PV output power. Traditionally, UI took a conservative approach by assuming the worst case scenario of zero customer load with PV generation at the nameplate rating. However, PV output is significantly affected by a range of meteorological variables that can vary widely across different geographical locations and time intervals, causing a PV system's peak power generation to deviate from its nameplate rating [7]–[9]. Since May 2015, approximately 250 PV interconnection applications failed to pass UI's screening process due to concerns with overvoltage. As more interconnection requests are received in the future, more applications will fail the screening process. It is, therefore, of great importance to obtain more accurate extreme PV output data.

Moreover, since PV power systems and solar-powered microgrids are increasingly integrated with utility grids, understanding extreme PV power characteristics has become critically important for ensuring grid resilience, security, reliability and beyond. Only a few critical applications are highlighted to demonstrate the importance of evaluating extreme PV power as an exhaustive survey for the applications of extreme PV analytics could be a paper in itself.

- Grid resilience. High PV power penetration causes dramatic midday dips of electricity demand, i.e., the so-called 'duck curves', which becomes a common phenomenon across the continent, from CAISO to ISO New England. The challenges to grid resilience include high system voltages and frequencies, unscheduled power flows into neighboring regions, and urgent need of power sources with fast ramp-up capabilities [10]. Accurate assessment of extreme generation of PV and other distributed energy resources (DERs) is the key to the development of resilience measures such as demand response, variable pricing, energy storage, and real-time fast-start pricing [10], just to name a few.
- Grid security. Extreme events, combined with the peaks of power generation and load, define grid security. Accurate models of extreme PV generation as well as other extremes in the grid will lead to accurate identification and mitigation of catastrophic contingencies and cascading events. Models for extreme PV and other grid variables are able to analyze the exceedances of power system limits, which are unattainable by existing forecasting and statistical planning tools that only look at the averaging effect of the distribution body and, therefore, cannot characterize the spatiotemporal features of the extremes.
- Grid reliability. The existing grid reliability tools, e.g., the Monte Carlo simulation, are incapable of forecasting extreme-event induced grid shifts because the sample size of a Monte Carlo simulation has to be extremely small to capture rare events. Focusing on the tails of events, the models for extreme events enable ultra-fast Monte Carlo simulations as the sample size can be significantly increased. This allows us to obtain reliability indices such as the System Average Interruption Duration Exceeding Threshold (SAIDET), the System Average Interruption Frequency Exceeding Threshold (SAIFET), and mean restoration times. This includes calculating probability distributions, mean and median values, deviations, skewness, and Kurtosis in an extremely fast manner.
- Other benefits, such as the increase of hosting capacity of PVs, protection of PV impacts on the grids, and improving the none-detection-zone prediction of PV integration [11], will not be discussed in detail due to limited space.

Extreme PV power, i.e., extremely high PV power, is an unusual value, thus, the data of extreme PV power at a given time interval are scarce. Extreme value analysis (EVA), which was pioneered by Fisher and Tippett [12], provides a promising approach for evaluating extreme phenomena in engineered systems. It uses a suitable probability distribution to fit a series of extreme data. Although EVA has been widely adopted in such areas as floods [13], [14], droughts [15], rainstorms [16], [17], and high winds [18], it has not yet been used to estimate extreme PV power. Part of the reason for this stems from the fact that the existing version of EVA cannot be directly applied to evaluating extreme PV power, because it requires multiple years of data. For instance, [13] uses 30 years of data from a subtropical

region of eastern Australia to evaluate extreme flood events, and the authors claim that even 30-years is not long enough for an accurate estimate. [17] analyzes extreme annual and seasonal rainfall patterns for 14 stations in Shaanxi, China, using 55 to 60 years of data collected from the 1950s to 2014. [19] utilizes 51 years of data spanning from 1941 to 1991 to conduct an EVA of icing on power lines. In other work, such as [14]–[16], [18], [20], [21], the recorded length of data collection spans from 18 to 120 years. Since most PV systems are recently installed and can provide only one or two years of data (or less), applying the traditional version of EVA to data from an individual PV system will not yield an accurate estimate.

To tackle this challenge, this paper presents Extreme PV Power Analytics (EPVA), a methodology that combines data from multiple PV systems for the purpose of an EVA. However, different PV systems may have different extreme PV power behaviors, making it challenging to determine which PV systems should be combined. To address this issue, the concept of the extreme capacity factor (ECF) is introduced, and a zone partitioning method, k-means clustering, is developed to divide the utility service territory into k clusters, such that PV systems in the same cluster will have similar ECF behaviors. EVA is subsequently applied to obtain the probability distribution of ECFs in each individual cluster. The main advantages are:

- EPVA is able to evaluate extreme PV power in any region and at any time interval, whereas the existing literature either analyzes normal power instead of extreme power [22]–[25] or fails to consider the spatiotemporal effect [26].
- A k-means clustering approach is developed to effectively cluster those PV systems that have similar ECF behaviors, which makes EVA applicable to real-life power systems where extreme PV data are usually scarce.
- A systematic approach is devised to obtain not only the distributions of PV ECFs but also the return levels and return periods. Future return intervals, which are hardly attainable through existing methods, are particularly important for system planning and risk analysis.
- One year of PV and weather data were collected from the UI service territory. Together with the EPVA results and insights obtained, the data will offer valuable resources for research communities and the power industry.

The rest of this paper is organized as follows: Section II is devoted to describing the data used in this paper, as well as the challenges encountered. Section III describes the EPVA approach. The zone partitioning results for the UI's territory are provided in Section IV, and the extreme PV capacity factor distributions for the UI's territory are presented in Section V. Section VI concludes the paper.

II. Data and Challenges Description

This section describes the data collected for the purposes of conducting EPVA. As an example, this study was conducted across 90 PV sites, each of which was sampled every 15 minutes throughout 2016, thus providing one year of output power data. Five meteorological variables from 9 weather sites were generated by the North American Mesoscale Forecast

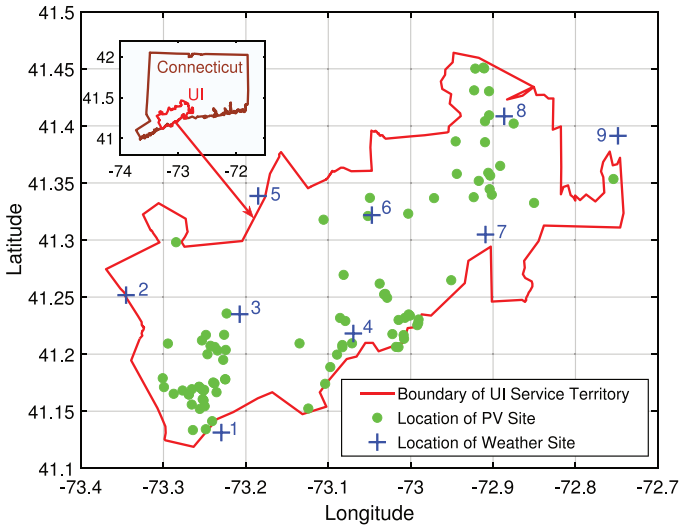


Fig. 1. Distribution of 90 PV and nine weather sites in UI service territory.

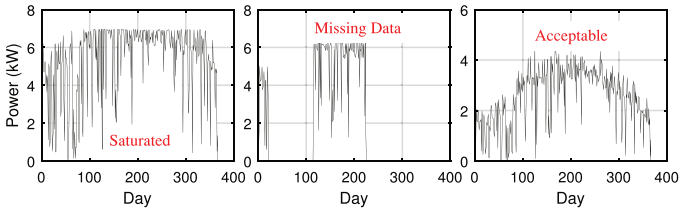


Fig. 2. The daily maximum PV power from three typical PV sites in 2016.

System (NAM) analysis: solar irradiance (W/m^2), 2-m humidity (%), 2-m temperature (K), 10-m wind speed (m/s), and cloud coverage (%) [27], [28]. Each meteorological variable was generated every three hours, and data were obtained from the National Centers for Environmental Prediction (NCEP) over 2016. The distribution of PV and weather sites is illustrated in Fig. 1. The details of the data include:

- Each PV site's information provides the site ID, latitude, longitude, nameplate rating (kW), and output power (kW) every 15 minutes throughout 2016.
- Each weather site's information provides the weather ID, latitude, longitude, and values of the five meteorological variables. The time resolution of each meteorological variable is three hours, and data were obtained over 2016.

Since PV output is greatly influenced by weather conditions, this paper aims to use the five meteorological variables listed above to divide the utility service territory into several groups at a given time interval, such that the PV systems in the same group have similar behaviors with regard to extreme output. However, it is challenging to achieve this objective. The main issues are as follows:

- Saturated or missing data. Fig. 2 displays the daily maximum PV output power for three typical PV sites. Each PV site in Fig. 2 provided one year of data throughout 2016, and the maximum power on each day is represented. Note that saturated and incomplete data are still useful and are, therefore, not filtered out.

- Spatiotemporally evaluating extreme PV power. PV output is significantly influenced by weather conditions, which exhibit great dissimilarities in different geographical locations and at different time intervals.
- Properly utilizing weather data to represent extreme PV output behaviors. PV and weather sites have different geographical locations, making it challenging to associate PVs with relevant weather data. Further, different meteorological variables have different impacts on PV output.

III. Extreme PV Power Analytics

In this section, the Extreme PV Power Analytics (EPVA) method is presented. It further develops k-means clustering to determine which PV sites have similar behaviors in PV extreme capacity factors (ECFs) and utilizes EVA to obtain the distributions of ECFs. The theoretical novelties of EPVA include:

- The concept of ECF is presented to better characterize PV output performance for different PV sites.
- Dividing the utility service territory into several clusters at a given time interval such that PV systems in the same cluster have similar behaviors in terms of ECFs is defined in k-means clustering, and once achieved, guarantees an accurate estimate of EVA.
- To achieve the objective in k-means clustering, various techniques are introduced, including: 1) considering the weight of each meteorological variable, 2) normalizing the meteorological variables, 3) determining representative data for each weather site, and 4) determining the optimal value of k.

The EPVA method developed in this paper aims to achieve a specific objective, considering various real challenges. This paper explicitly addresses the challenges in the process and provides guidance that is easy to follow for electric utilities when obtaining accurate performance data on extreme PV power in any region and at any time interval.

A. Extreme Capacity Factor

PV power systems have different nameplate ratings, which lead to different peak outputs. In order to better characterize PV output performance for different PV sites, this paper presents the concept of capacity factor (CF), as defined below, to normalize PV output power:

$$C_{ft} = \frac{P_t}{P_r}, \tag{1}$$

where C_{ft} is PV CF at time t, while P_t and P_r are the output AC power at time t and the AC nameplate rating, respectively.

It should be noted that the nameplate rating in a PV system is reported in terms of the aggregated capacity of either all modules (with DC nameplate ratings), or all inverters (with AC nameplate ratings). The conversion between AC and DC nameplate ratings can be achieved by the ratio between the output power of the DC solar array and the AC inverters.

In order to evaluate a system's extreme power performance, the concept of ECF is defined in the following way:

$$E_{cf\Delta t} = M_{ax}(C_{f1}, C_{f2}, \ldots, C_{fn}), \tag{2}$$

where C_{f1}, C_{f2}, ..., C_{fn} are CFs of one PV site at time t_1, $t_2, \ldots, t_n$, respectively. $M_{ax}()$ represents a function that selects the maximum value from the contents. $E_{cf\Delta t}$ is the ECF of that PV site at the specific time interval from t_1 to t_n.

Note that if two PV systems have a long geographical distance, then the weather conditions are likely to exhibit great dissimilarities at a given time interval, thus, resulting in different values of ECF. But it is also likely that two PV systems with a large geographical distance have similar values of ECFs.

(2) shows that $E_{cf\Delta t}$ is only one value, namely the maximum value among CFs at a given time interval. In order to gather more data for a more accurate distribution, $E_{cf\Delta t}$ can be determined as a series of top values among CFs, and a threshold is used to obtain $E_{cf\Delta t}$. That is,

$$E_{cf\Delta t} = M_{ax}^{'}(C_{f1}, C_{f2}, \ldots, C_{fn}), \tag{3}$$

where $M_{ax}^{'}()$ is a function that selects the top values from the contents according to a given threshold.

B. K-Means Clustering

K-means clustering is developed in this paper to determine which PV systems have similar behaviors in their PV ECFs. A flow chart for this process is shown in Fig. 3. In doing this, more extreme data can be combined for the purposes of EVA. The following subsections present the details of properly using PV and meteorological data to achieve this objective.

1) Objective Function: K-means clustering aims to partition all the observations into k clusters $C = (C_1, C_2, \ldots, C_k)$, such that the observations in the same cluster have great similarity [29], [30]. The basic procedures of k-means clustering are 1) initially dividing all the observations into k clusters, 2) calculating the distance between each observation and the centroid of each cluster, 3) classifying each observation into its closest cluster, 4) redetermining the centroid of each cluster, and 5) repeating steps 2 through 4 until a convergence is reached. The objective function is defined as follows [31]:

$$V = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2, \tag{4}$$

where V is the within-cluster sum of squares (WCSS), and it will be minimized when convergence has been reached. x_j and μ_i represent the jth observation within cluster C_i and the centroid of C_i, respectively. $||x_j - \mu_i||$ is the Euclidean distance between x_j and μ_i.

Note that the determination of this distance is the key in k-means clustering, since the distance acts as an indicator of the degree to which two observations are similar. The smaller the distance is, the greater the similarity will be. In this study, as the behavior of PV output is largely affected by weather conditions, a range of meteorological variables are used to calculate the distance. By classifying weather sites into k clusters, PV sites, whose closest weather sites belong to the same cluster, will have similar behaviors in their extreme PV outputs.

Different meteorological variables have different degrees of influence on PV output power, and therefore the weight of

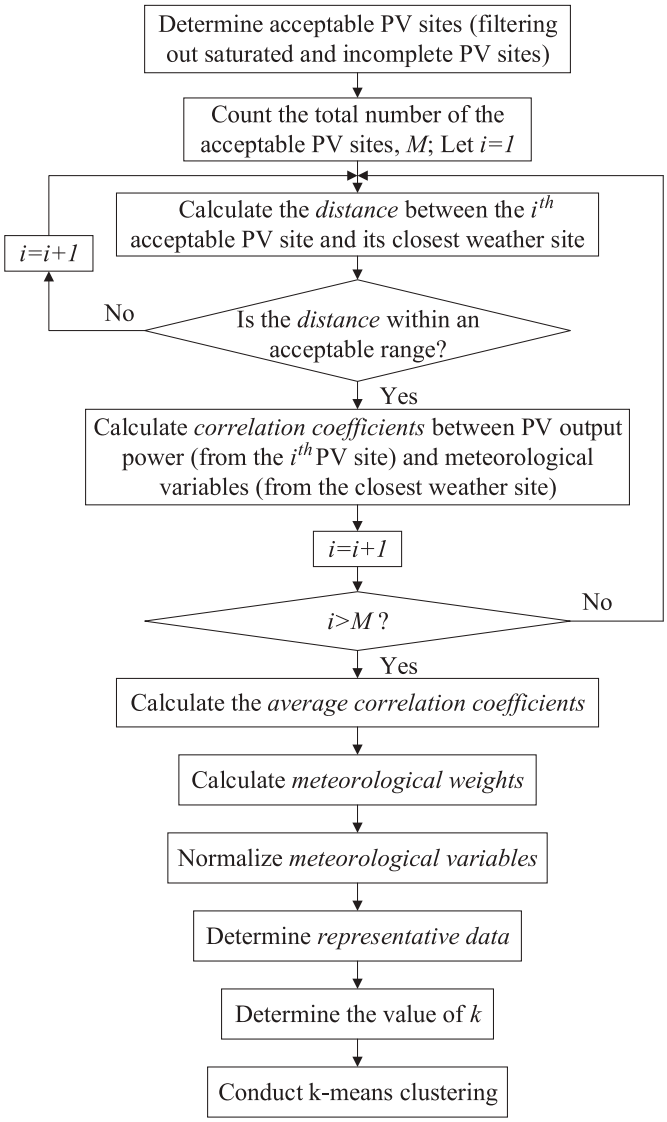


Fig. 3. The flowchart of conducting k-means clustering in EPVA.

each meteorological variable, which reflects the degree of influence, should be considered in the calculation of Euclidean distance. Considering the weights of meteorological variables, $||x_j - \mu_i||^2$ in (4) can be expressed as

$$||x_j - \mu_i||^2 = W_s^2(x_{js} - \mu_{is})^2 + W_h^2(x_{jh} - \mu_{ih})^2 + W_t^2(x_{jt} - \mu_{it})^2 + W_w^2(x_{jw} - \mu_{iw})^2 + W_c^2(x_{jc} - \mu_{ic})^2, \quad (5)$$

where x_{js}, x_{jh}, x_{jt}, x_{jw} and x_{jc} are the five meteorological variables of the jth weather site within cluster C_i, namely, solar irradiance, humidity, temperature, wind speed and cloud coverage, respectively. μ_{is}, μ_{ih}, μ_{it}, μ_{iw} and μ_{ic} are the five meteorological variables of the centroid of C_i, respectively.

Note that although (5) does not consist of PV power, the meteorological variables in this equation are actually selected based on the extreme PV output power, which will be discussed later.

2) Calculate Meteorological Weights: The weight of each meteorological variable is associated with the correlation coefficient between PV output power and each meteorological variable. The larger the absolute value of the correlation coefficient is, the higher the weight will be. A common method for calculating the correlation coefficient, the Pearson product-moment correlation, is defined as follows [32]:

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}, \quad (6)$$

where r is the correlation coefficient, N is the number of total data points, X_i and Y_i are PV power and one meteorological variable of the ith data point, respectively. $\bar{X}$ and $\bar{Y}$ are the averages of X_i and Y_i in the N-data set, respectively.

Before calculating correlation coefficients, three concerns should be addressed: 1) PV sites with saturated or missing data should be removed,[1] otherwise the calculated correlation coefficients will not reflect the actual relationship between PV output and meteorological variables; 2) each PV site should be associated with its closest weather site, such that the meteorological variables generated from that weather site will share the most similar weather conditions with its PV site; and 3) if the distance between one PV site and its closest weather site is still large, that pair of PV and weather sites should not be considered in the calculation of correlation coefficients.

It is likely that the same weather site is associated with multiple PV sites, which means these PV sites share the same weather conditions.

(6) gives the calculation of the correlation coefficient for one PV site. To achieve an accurate result, the average is calculated from multiple PV sites as follows:

$$\bar{r} = \frac{1}{N_1}\sum_{i=1}^{N_1} r_i, \quad (7)$$

where $\bar{r}$ is the average value of the correlation coefficient, N_1 is the number of the PV sites available for calculating correlation coefficients, and r_i is the correlation coefficient from the ith PV site.

Meteorological weights are then calculated according to their ratios of correlation coefficients. For instance, the weight for solar irradiance is calculated as follows:

$$W_s = \frac{|\bar{r}_s|}{|\bar{r}_s| + |\bar{r}_h| + |\bar{r}_t| + |\bar{r}_w| + |\bar{r}_c|}, \quad (8)$$

where W_s is the weight of solar irradiance. $\bar{r}_s$, $\bar{r}_h$, $\bar{r}_t$, $\bar{r}_w$ and $\bar{r}_c$ are the average correlation coefficients of the five meteorological variables, namely, solar irradiance, humidity, temperature, wind speed and cloud coverage, respectively. $|\cdot|$ represents the absolute value of the content. The weights of humidity, temperature, wind speed and cloud coverage can be calculated in a similar way, and are represented as W_h, W_t, W_w and W_c, respectively.

[1]Note: Saturated and incomplete data are still useful in extreme value analysis (EVA); they are just filtered out in the k-means clustering stage.

3) Normalize Meteorological Variables: Each meteorological variable should be normalized, because different meteorological variables have different values due to differences in their units. Using solar irradiance as an example, x_{js} in (5) is normalized as follows:

$$x_{jsn} = \frac{x_{js} - x_{si}}{x_{sa} - x_{si}}, \tag{9}$$

where x_{jsn} is the normalized value of x_{js}. x_{si} and x_{sa} are the minimum and maximum solar irradiance among the data from all weather sites within a given time interval.

4) Determine Representative Data: Another key but challenging component in the calculation of Euclidean distance is determining what constitutes representative data. According to (5), for one weather site such as x_j, each meteorological variable (x_{js}, x_{jh}, x_{jt}, x_{jw}, or x_{jc}) should only have one value, such that the distance between each weather site and any centroid can be calculated. However, every weather site has a series of data (generated every three hours in this study) within a given time interval. In other words, among a series of data, only one data will be used in the distance calculation, and the data should represent for the weather site in terms of extreme PV power behaviors.

In the presented EPVA, for one weather site at a given time interval, the groups of meteorological variables that relate to the top 1% of PV power (from the closest PV plant) are selected. For instance, each weather site has 248 groups of meteorological variables in January, because the time resolution of weather data is three hours. Only two groups of meteorological variables relate to the top 1% of PV power, and these two groups of meteorological variables will be selected. Note that 1) each group of meteorological variables contains solar irradiance, humidity, temperature, wind speed and cloud coverage at one time moment; and 2) PV power will convert its time resolution from 15 minutes to three hours in this stage to fit weather data. Each meteorological variable will then be averaged from the selected groups to become the representative data.

Mathematically, suppose for the weather site x_j, there are N_2 groups of meteorological variables within a specific time interval that relate to the top 1% of PV power. The selected representative solar irradiance for x_j is calculated as follows:

$$X_{js} = \frac{1}{N_2} \sum_{i=1}^{N_2} x_{jsni}, \tag{10}$$

where X_{js} is the representative solar irradiance for x_j, and x_{jsni} is the ith normalized solar irradiance of x_j.

Similarly, X_{jh}, X_{jt}, X_{jw} and X_{jc}, which represent the selected humidity, temperature, wind speed and cloud coverage for x_j, can also be calculated in the same way as X_{js}. Therefore, (5) can be reexpressed as

$$||x_j - \mu_i||^2 = W_s^2(X_{js} - \mu_{is})^2 + W_h^2(X_{jh} - \mu_{ih})^2 + W_t^2(X_{jt} - \mu_{it})^2 + W_w^2(X_{jw} - \mu_{iw})^2 + W_c^2(X_{jc} - \mu_{ic})^2. \tag{11}$$

5) Determine k: The Silhouette value reflects how well a weather site belongs to its cluster [33]. The higher the Silhouette value is, the more accurate the clustering result for that weather site will be. Therefore, the average Silhouette value for N_3 weather sites can be calculated to evaluate how well the N_3 weather sites are clustered, as shown below:

$$\bar{S} = \frac{1}{N_3} \sum_{i=1}^{N_3} S_j, \tag{12}$$

where $\bar{S}$ is the average Silhouette value, and S_j is the Silhouette value of weather site x_j, which is calculated by

$$S_j = \frac{b_j - d_j}{M_{ax}(b_j, d_j)}, \tag{13}$$

where b_j is the lowest average distance between x_j and the other weather sites in any other cluster, and d_j is the average distance between x_j and other weather sites in the same cluster. Suppose x_j belongs to the cluster C_1. Then, b_j and d_j are calculated as follows:

$$b_j = M_{in}\left(\frac{1}{n_2}\sum_{x_{i_2}\in C_2}||x_j - x_{i_2}||, \frac{1}{n_3}\sum_{x_{i_3}\in C_3}||x_j - x_{i_3}||, \cdots, \frac{1}{n_k}\sum_{x_{i_k}\in C_k}||x_j - x_{i_k}||\right) \tag{14}$$

and

$$d_j = \begin{cases} \frac{1}{n_1 - 1}\sum_{x_{i_1}\in C_1/x_j}||x_j - x_{i_1}||, & n_1 > 1 \\ 0, & n_1 = 1 \end{cases} \tag{15}$$

where $n_1, n_2, \ldots, n_k$ are the number of weather sites in clusters $C_1, C_2, \ldots, C_k$, respectively ($n_1 + n_2 + \cdots + n_k = N_3$). $M_{in}()$ is a function that selects the minimum value from its contents.

Because a successful clustering usually has a high $\bar{S}$, in the presented EPVA k is determined by selecting the number that has the maximum $\bar{S}$.

C. Extreme Value Analysis

In this section, the extreme value analysis (EVA) of PV ECFs is described, including 1) a description of the probability distributions and 2) a calculation of the return level.

1) Probability Distributions: EVA aims to use a specific distribution (e.g., the generalized logistic, the generalized extreme value, the log-normal, or the generalized Pareto distribution) to model the tail of another distribution [34]. The generalized Pareto distribution (GPD) is a frequently-used distribution for various applications, such as rainfall [35], high wind [36], false data in open source software [37], and inverse synthetic aperture radar imaging [38]. This paper will use GPD to illustrate how to evaluate the distribution of PV ECFs and validate the superiority of EPVA.

GPD is specified by three parameters: u (location), α (scale) and γ (shape), and is described as follows:

$$P(X > x|X > u) = [1 + \gamma(x - u)/\alpha]^{-1/\gamma}, \tag{16}$$

where $P(X > x|X > u)$ is the probability of exceeding x given the condition that X is above u. The three parameters can be

estimated through the combination of L-moments with extreme data. The details can be found in [39].

The probability density function (PDF), $f_{u,\alpha,k}(x)$, can be calculated with the three parameters, u, α and k, as follows:

$$f_{u,\alpha,\gamma}(x) = \frac{1}{\alpha}\left(1 + \frac{\gamma(x-u)}{\alpha}\right)^{\left(-\frac{1}{\gamma}-1\right)}. \quad (17)$$

The PDF is widely used to specify the probability within a given interval. For instance, the probability of X falling between the values a and b can be calculated as

$$P(a \le X \le b) = \int_a^b f_{u,\alpha,\gamma}(x)\,\mathrm{d}x. \quad (18)$$

2) Return Level: Return level and return period are commonly used in EVA to help estimate the future return interval, which can be useful for system planning and risk analysis. A return level with a return period of $1/p$ years is a threshold whose probability of being exceeded is p [40]. From (16), the probability of exceeding x is

$$P(X > x) = \zeta_u[1 + \gamma(x-u)/\alpha]^{-1/\gamma}, \quad (19)$$

where $\zeta_u = P(X > u)$ is the probability of exceeding u, and can be estimated as follows:

$$\zeta_u = \frac{m_u}{m_s}, \quad (20)$$

where m_s is the number of the total samples and m_u is the number of the samples exceeding u. According to (19), the level x_r that is exceeded on average once every r observations is the solution of

$$\frac{1}{r} = \zeta_u[1 + \gamma(x_r - u)/\alpha]^{-1/\gamma}. \quad (21)$$

By rearranging (21), x_r is expressed as

$$x_r = u + \frac{\alpha}{\gamma}[(r\zeta_u)^{\gamma} - 1]. \quad (22)$$

Let $r = N \times n_y$, where N is the number of years and n_y is the number of samples each year. The N-year return level, a value expected to be exceeded once every N years, is

$$z_N = u + \frac{\alpha}{\gamma}[(Nn_y\zeta_u)^{\gamma} - 1]. \quad (23)$$

IV. Case Exemplar for Zone Partitioning

In this section, the zone partitioning results for the UI's territory are presented based on k-means clustering (conducted in Matlab), which include 1) acceptable PV sites, 2) the closest weather site for each PV site, 3) correlation coefficients, 4) meteorological weights, 5) normalized meteorological variables, 6) values of k for different months, and 7) clustering results.

A. Acceptable PV Sites

The daily maximum power for all 90 PV sites was obtained by the same process shown in Fig. 2, and was then scrubbed to eliminate the PV sites with saturated or missing data. This reduces the number of PV sites down to 34. The IDs of the acceptable, saturated and incomplete PV sites are given in Table I. It can be seen that, among the 90 PV sites, the number of saturated PV sites is more than that of the acceptable PV sites, which is a result of the strict interconnection requirement described in the Introduction.

TABLE I
The Acceptable, Saturated and Incomplete PV Sites

Acceptable PV ID:						
911003	1725003	3366733	3380628	3381784	3382270	3383779
3384607	3385321	3385395	3386582	3457134	3583066	3129001
3381050	3382785	3386589	3394947	3364558	3420277	3424695
3437325	3434324	2936001	3426832	3429680	3500077	3504991
3361693	3375507	3378557	3379947	3414574	3451501	
Saturated PV ID:						
3440308	3420288	3420283	3425582	3423150	3457666	3507463
3392301	3393215	3539217	3429609	1902001	3434276	3437568
3419058	3386583	3384516	3385032	3464492	3408547	3429688
3424697	3376313	2530012	3429453	3391952	3485054	3410606
3381938	3424706	3437544	3422429	3389745	3380733	3422404
3400108	3383582	3429165	3381167	3318022	3393207	3385396
3525066	3386578	3380923	3505985	3484753	3417764	3399938
3431331	3429188	3499091				
Incomplete PV ID:						
3437520	3381579	3362732	3424697			

TABLE II
The Closest Weather Site for Each PV Site

PV	911003	1725003	3366733	3380628	3381784	3382270
Weather	1	1	1	1	1	1
D (km)	4.46	3.2	3.85	4.53	5.38	4.88
PV	3383779	3384607	3385321	3385395	3386582	3457134
Weather	1	1	1	1	1	1
D (km)	3.47	4.46	8.35	1.99	5.43	3.73
PV	3583066	3129001	3381050	3382785	3386589	3394947
Weather	1	2	3	3	3	3
D (km)	5.37	6.89	4.66	2.72	4.34	5.34
PV	3364558	3420277	3424695	3437325	3434324	2936001
Weather	4	4	4	4	6	7
D (km)	6.42	4.23	2.82	8.34	6.12	3.72
PV	3426832	3429680	3500077	3504991	3361693	3375507
Weather	7	7	7	7	8	8
D (km)	3.73	6.61	6.01	5.40	6.55	2.47
PV	3378557	3379947	3414574	3451501		
Weather	8	8	8	8		
D (km)	2.99	4.48	5.68	3.38		

B. Closest Weather Site for Each PV Site

The results of the closest weather site for each acceptable PV site are presented in Table II. It can be seen that 1) different PV sites have different distances from their closest weather sites, and 2) the range of distances between most PV sites and their closest weather sites is from 1.99 km to 6.89 km, with the exception of two PV sites whose closest distances are larger (8.35 km and 8.34 km, respectively). These two PV sites can thus be eliminated from the calculation of correlation coefficients.

C. Correlation Coefficients

Fig. 4(a) displays the absolute values of the calculated correlation coefficients (between PV output power and each meteorological variable) from the selected 32 PV sites in January. It shows that, different meteorological variables have different levels of absolute correlation coefficients. Solar irradiance

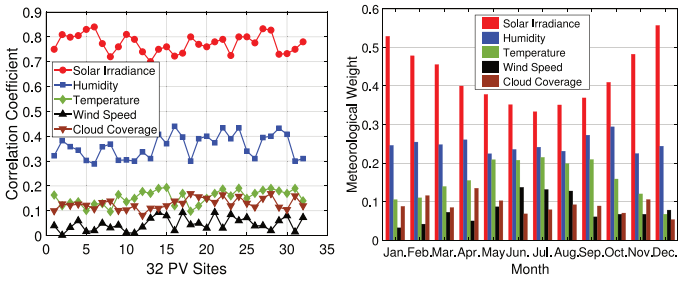


Fig. 4. Results of correlation coefficients and meteorological weights. (a) Absolute values of correlation coefficients in January. (b) Meteorological weights in different months.

has the largest value and is followed by humidity, whereas the other three meteorological variables have relatively small values.

This result is reasonable. Correlation coefficients reflect how strong a relationship is between two variables. Solar irradiance has the largest correlation coefficient, because solar cells are made of semiconductor materials, and when light strikes the solar cells, electrons are knocked loose from the atoms and can be captured in the form of an electric current. The stronger the solar irradiance is, the larger the current/power will be. Humidity also plays an important role in affecting PV output power. It brings down the utilization of solar energy, and reduces output power from solar panels [41]–[43]. There are two ways that humidity affects output power. First, water vapour particles will be refracted, reflected or diffracted when light hits water droplets. Second, humidity ingression to the solar cell enclosure degrades the performance of solar cells [43]. Temperature negatively affects solar cell performance. When temperature increases, the concentration of internal carriers increases, and the internal carrier recombination rates also increase [44]. Wind speed has a negative effect on relative humidity, which indirectly affects the received solar irradiance. However, the correlation of wind speed is much smaller than that of humidity. Cloud coverage has a low correlation coefficient in this study. The reason is, in the meteorological database the cloud coverage data were recorded not only in the daytime, but also at night. At night, PV output power is zero, but the cloud coverage would still vary continuously and oftentimes significantly. Only the daytime cloud coverage has an impact on PV power by affecting solar irradiance, whereas it does not have an impact on PV output at night, no matter how variable it can be.

Fig. 4(a) also illustrates that, for different PV sites, the correlation coefficients of each meteorological variable have the same levels of values. It is, therefore, reasonable to average the correlation coefficients of each meteorological variable from all the PV sites (refer to (7)). Using January as an example, the average correlation coefficients are provided in Table III. Note that 1) the larger the absolute value of correlation coefficient is, the greater the effect of the meteorological variable on the PV output power will be; and 2) although humidity and cloud coverage have negative correlation coefficients, they

TABLE III
AVERAGE CORRELATION COEFFICIENTS IN JANUARY

Solar Irradiance	Humidity	Temperature	Wind Speed	Cloud Coverage
0.7733	-0.3596	0.1540	0.0468	-0.1284

TABLE IV
MAXIMUM AND MINIMUM VALUES OF DIFFERENT METEOROLOGICAL VARIABLES IN DIFFERENT MONTHS

	x_{sa} W/m^2	x_{si} W/m^2	x_{ha} %	x_{hi} %	x_{ta} K	x_{ti} K	x_{wa} m/s	x_{wi} m/s	x_{ca} %	x_{ci} %
Jan.	514	0	99	24	287	259	12.26	0.06	100	0
Feb.	675	0	100	33	288	251	11.63	0.10	100	0
Mar.	848	0	100	26	298	267	11.28	0.10	100	0
Apr.	915	0	100	17	300	268	12.37	0.04	100	0
May	958	0	100	26	306	277	9.16	0.01	100	0
Jun.	969	0	100	27	305	281	9.21	0.01	100	0
Jul.	951	0	100	32	309	286	7.24	0.07	100	0
Aug.	914	0	100	34	309	285	8.62	0.03	100	0
Sep.	808	0	100	32	306	277	10.38	0.06	100	0
Oct.	659	0	100	32	302	273	11.94	0.07	100	0
Nov.	513	0	100	28	294	271	12.13	0.04	100	0
Dec.	408	0	100	29	288	261	12.00	0.14	100	0

are still considered in the calculation of Euclidean distance in (5), because the purpose of calculating the Euclidean distance is to find the dissimilarity between two weather sites.

D. Meteorological Weights

The results of meteorological weights in different months are illustrated in Fig. 4(b). It can be seen that 1) the weights of different meteorological variables are different. Solar irradiance has the largest weight in any month, followed by humidity. The other three meteorological variables, namely temperature, wind speed and cloud coverage, have relatively small weights. 2) At different time intervals, the weights of each meteorological variable are also different. For instance, the weight of solar irradiance in summer is lower than it is in winter, while the temperature's weight has the opposite tend.

E. Normalized Meteorological Variables

Table IV provides the maximum and minimum values of the five meteorological variables. It shows that 1) the meteorological variables have different levels of values, and that's why they should be normalized for better comparison in the calculation of Euclidean distance (refer to (11)), and 2) some of the meteorological variables vary widely at different time intervals, indicating that, for different time intervals, different maximum and minimum values should be utilized in the normalization of meteorological variables. With the maximum and minimum values, normalized meteorological variables are calculated according to (9).

F. The Values of k

The selected k and their corresponding average Silhouette values, $\bar{S}$, for different months are shown in Table V. It can be seen that 1) the optimal value of k is different at different

TABLE V
THE SELECTED k IN DIFFERENT MONTHS

	Jan.	Feb.	Mar.	Apr.	May	Jun.
k	4	4	5	4	3	5
$\bar{S}$	0.5914	0.5964	0.6416	0.50223	0.6750	0.5667
	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
k	5	3	5	5	4	3
$\bar{S}$	0.5118	0.5858	0.6022	0.5760	0.5246	0.5490

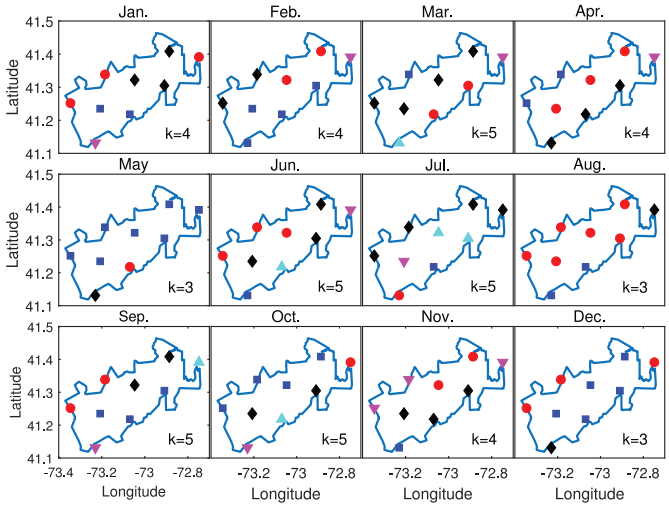


Fig. 5. Zone partitioning results for different months.

time intervals, and 2) with the selected k, the average Silhouette values are all above 0.5, indicating successful clustering results (Silhouette value ranges from -1 to 1. The larger the Silhouette value, the better the clustering result).

G. Clustering Results

Fig. 5 illustrates the zone partitioning results for the UI's territory in different months. In each subplot, the weather sites' different colors/shapes represent different clusters. It can be seen that 1) at different time intervals, the value of k can be different, and 2) even with the same k, the clustering results can also be different. Note that each PV system can be classified into its closest weather site based on the geographical distances. PV systems within the same cluster will have similar weather conditions in terms of ECFs.

V. EPVA TEST RESULTS FOR UI

In this section, part of the EPVA results obtained from R tools are presented, which include 1) probability density function (PDF) results in January, July, spring, and autumn in 2016; 2) their corresponding return level results; and 3) comparison between EPVA and local EVA results.

A. PDF Results

Figs. 6–9 give the PDF results for different clusters in January, July, spring, and autumn in 2016, respectively. The red curve in each subplot represents the modeled PDF based on the sample

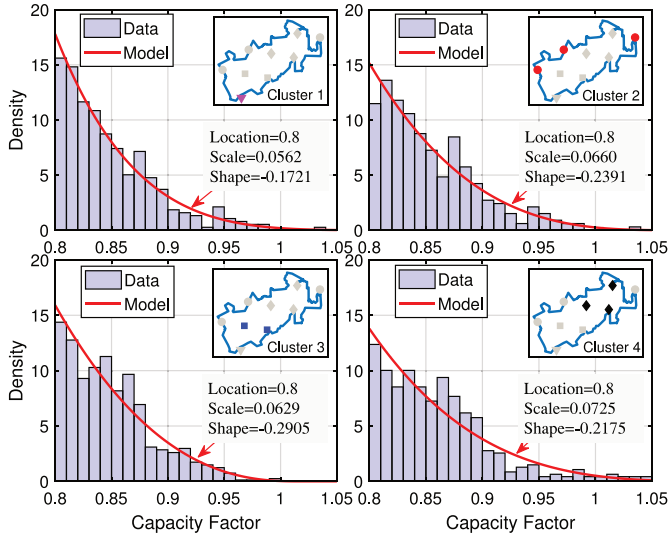


Fig. 6. Probability density function of PV ECF for each cluster in January.

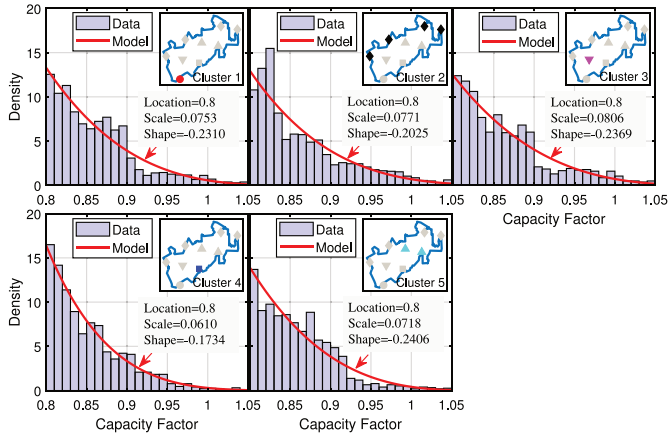


Fig. 7. Probability density function of PV ECF for each cluster in July.

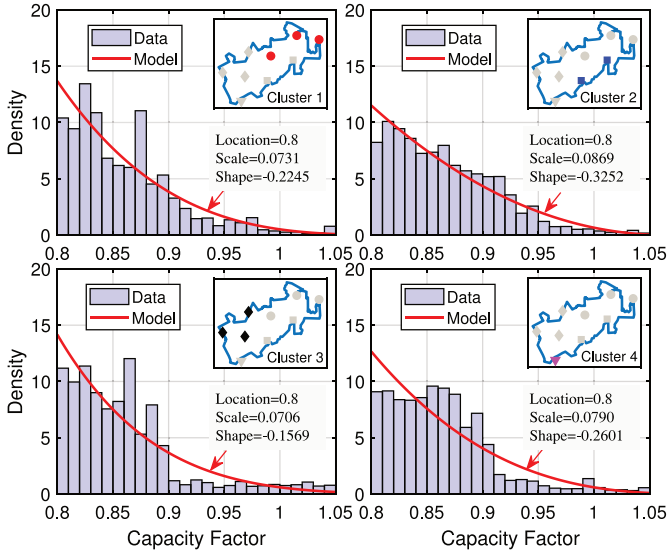


Fig. 8. Probability density function of PV ECF for each cluster in spring.

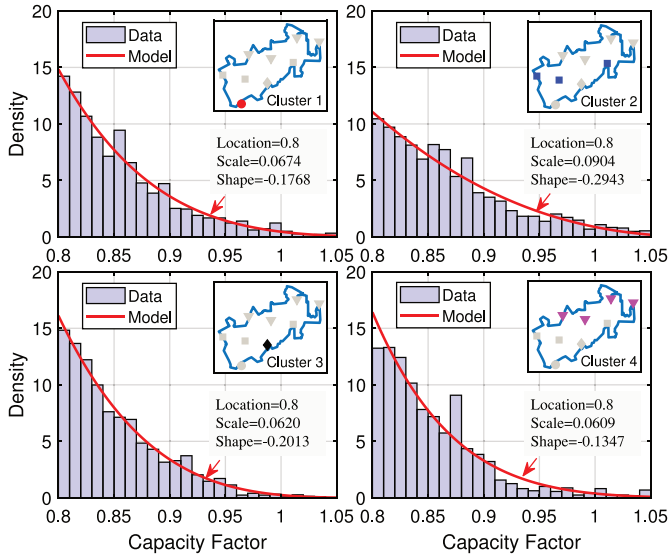


Fig. 9. Probability density function of PV ECF for each cluster in autumn.

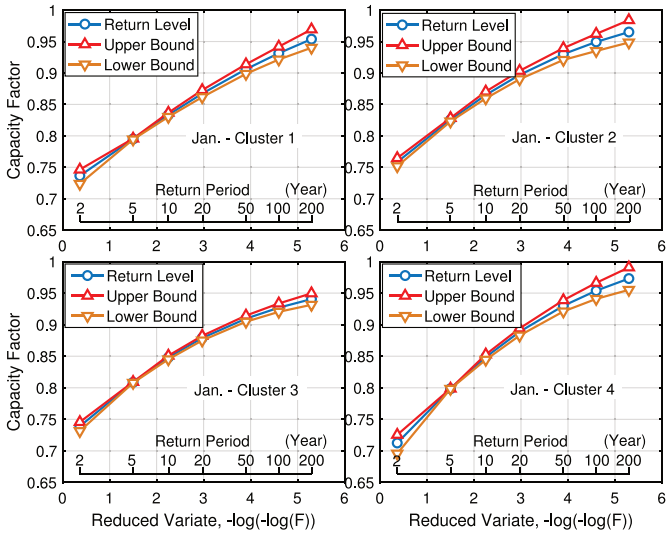


Fig. 10. Return levels, with their 90% error bounds for each cluster in January.

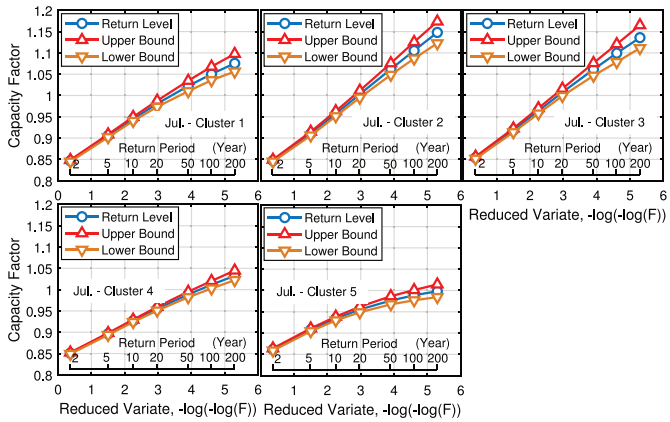


Fig. 11. Return levels, with their 90% error bounds for each cluster in July.

data from multiple PV systems in the same cluster at a given time interval. The three parameters of each PDF are also given. The location, u, is fixed at 0.8, indicating that the values above 0.8 are selected as the extreme values in this study. The x axis represents the capacity factor, while the y axis is the probability density. The probability of a capacity factor that falls between any two values can be calculated as the integration of the PDF within that interval.

From Figs. 6–9, it can be seen that 1) the modeled PDFs are well fit to the extreme data, which benefits from the large amount of data combined from multiple PV systems; 2) different clusters have different PDFs at a given time interval; and 3) at different time intervals, the PDFs are also different. Different PDFs lead to different distributions of return levels, as will be presented in the next subsection.

B. Return Level Results

Figs. 10–13 illustrate the results of return levels with their 90% error bounds for different clusters in January, July, spring, and autumn, respectively. In each subplot, the x axis denotes the return period, while the y axis is the capacity factor. The return period, F, is scaled through the function: -log(-log(F)). The width of 90% error bounds (upper bound - lower bound) can serve as a probabilistic indicator of the accuracy of the results, and the upper and lower bounds are obtained through Monte Carlo simulations (1000 runs). Note that the simulations should match the particular characteristics of the data such as the same number of PV sites and the same record length at each site. In general, error bounds illustrate the uncertainty of the return value, and should be used where a person is comfortable with the uncertainty but is not so strict. 90% error bounds are commonly used in extreme value analysis [45]–[47], and are therefore also used in this study. The larger the width of 90% error bounds, the less accurate the results are.

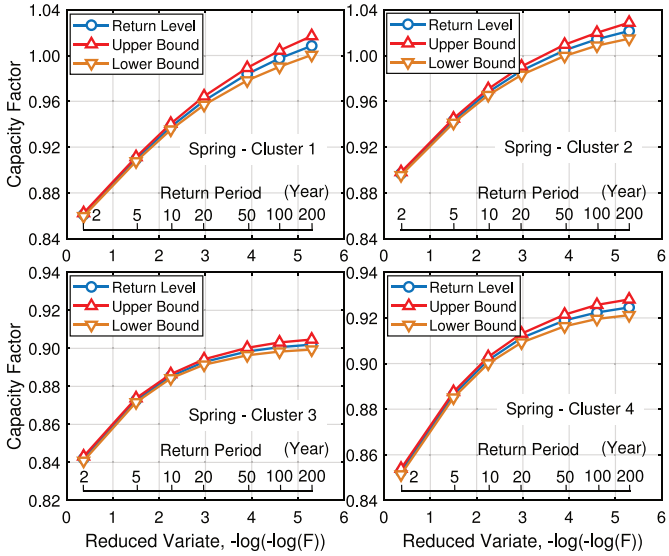


Fig. 12. Return levels, with their 90% error bounds for each cluster in spring.

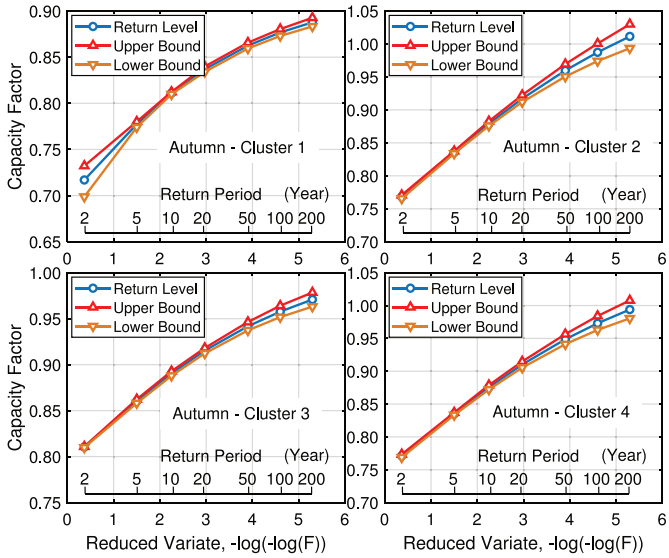


Fig. 13. Return levels, with their 90% error bounds for each cluster in autumn.

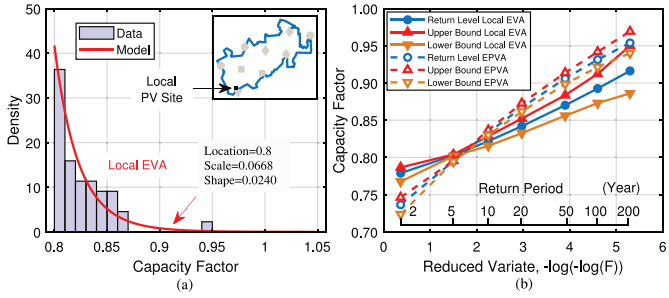


Fig. 14. Comparison of EPVA and local EVA results. (a) PDF of local EVA in January. (b) Return levels, with their 90% error bounds for EPVA and local EVA in January.

It can be seen that 1) the return level increases when the return period increases for any cluster at a given time interval. This is reasonable, because the higher the return level, the lower the probability of being exceeded; 2) the width of the 90% error bounds becomes larger when the return period is higher, indicating the results are less accurate; 3) different clusters have different return levels at a given time interval; and 4) at different time intervals, the return levels are also different. Compared with those in January, the return levels in July are much higher. This is reasonable, since the solar irradiance, which has the largest weight on PV output power (refer to Fig. 4), is stronger in summer compared with that in winter. The return levels in spring are at similar levels with those in autumn, as shown in Figs. 12 and 13.

The return levels obtained above offer valuable resources for research communities and the power industry in system planning and risk analysis.

C. Comparison of EPVA and Local EVA Results

The comparison results of EPVA and local EVA in January, July, spring, and autumn are shown in Figs. 14–17, respectively.

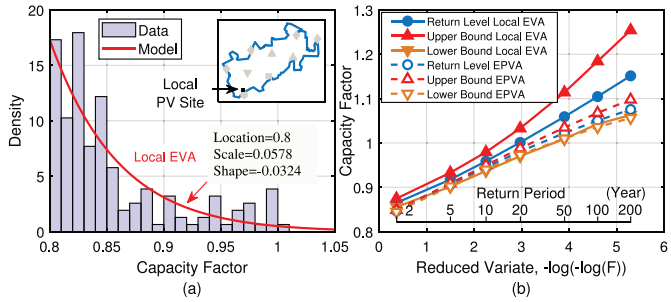


Fig. 15. Comparison of EPVA and local EVA results. (a) PDF of local EVA in July. (b) Return levels, with their 90% error bounds for EPVA and local EVA in July.

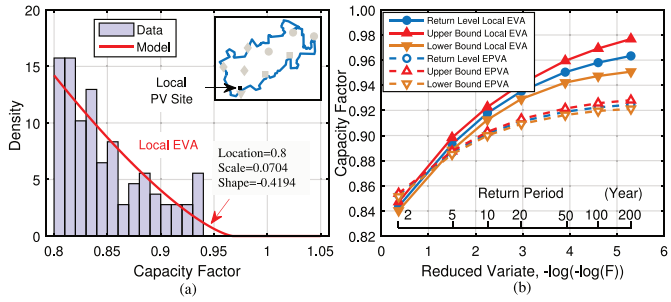


Fig. 16. Comparison of EPVA and local EVA results. (a) PDF of local EVA in spring. (b) Return levels, with their 90% error bounds for EPVA and local EVA in spring.

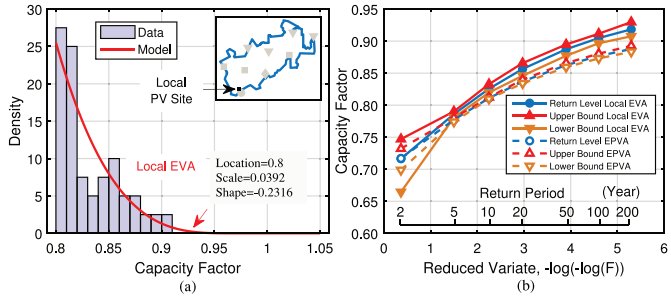


Fig. 17. Comparison of EPVA and local EVA results. (a) PDF of local EVA in autumn. (b) Return levels, with their 90% error bounds for EPVA and local EVA in autumn.

Figs. 14–17(a) illustrate the PDFs of local EVA at different time intervals based on the data from a local PV site. It can be seen that, compared with the results of EPVA, the modeled PDFs of a local EVA are not well fit to the data, which is mainly due to the less extreme data from an individual PV site. Figs. 14–17(b) give the comparison results of return levels in January, July, spring, and autumn, respectively. It can be seen that, compared with the local EVA, EPVA has a narrower width of the 90% error bounds, especially when the return period increases, indicating higher accuracy.

VI. Conclusion

This paper presents an EPVA approach to evaluating the extreme PV power in any region and at any time interval for electric utilities. K-means clustering is developed to divide the whole

territory into k clusters, such that PV systems in the same cluster have similar behaviors in their ECFs and can be combined to provide more extreme data for the purpose of EVA. Based on the combined data, EVA is subsequently used to obtain the distributions of PV ECFs. Compared with the local EVA, EPVA achieves more accurate estimates. As an outcome of this research, the method is to be further developed as a powerful tool for system planning, operation, and distributed energy resource integration. EVA in any region and at any time interval is particularly useful for reliability and resiliency monitoring and extreme power forecasting.

The remaining problems of this method include: 1) Each PV site and its closest weather site still have a certain distance, making the weather conditions from the weather site different from those at the PV site. This can be solved by increasing the resolution of weather sites. 2) For some utilities with small territories, k-means clustering may not be very useful, as the variation of weather conditions is small. In this case, there is no need to divide the territory into k clusters since all the PV sites within the territory have similar weather conditions, and thus similar ECF behaviors.

REFERENCES

[1] J. Zhang, H. Cho, R. Luck, and P. J. Mago, "Integrated photovoltaic and battery energy storage (PV-BES) systems: An analysis of existing financial incentive policies in the US," *Appl. Energy*, vol. 212, pp. 895–908, 2018.

[2] T. Cook, L. Shaver, and P. Arbaje, "Modeling constraints to distributed generation solar photovoltaic capacity installation in the US Midwest," *Appl. Energy*, vol. 210, pp. 1037–1050, 2018.

[3] D. Moskovkin, A. M. Mathew, Q. Guo, R. Eyetsemitan, and T. U. Daim, "Landscape analysis: Regulations, policies, and innovation in photovoltaic industry," in *Infrastructure and Technology Management*, Berlin, Germany: Springer, 2018, pp. 3–17.

[4] Y. Yang, Q. Ye, L. J. Tung, M. Greenleaf, and H. Li, "Integrated size and energy management design of battery storage to enhance grid integration of large-scale PV power plants," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 394–402, Jan. 2018.

[5] N. Kumar, I. Hussain, B. Singh, and B. K. Panigrahi, "Peak power detection of PS solar PV panel by using WPSCO," *IET Renewable Power Gener.*, vol. 11, no. 4, pp. 480–489, 2017.

[6] Public Utilities Regulatory Authority, "Regulations of Connecticut State Agencies," 2015. [Online]. Available: http://ct.gov/pura/lib/pura/regs/16-11-100to238.pdf

[7] J. Moshövel *et al.*, "Analysis of the maximal possible grid relief from PV-peak-power impacts by using storage systems for increased self-consumption," *Appl. Energy*, vol. 137, pp. 567–575, 2015.

[8] J. Enslin and D. Snyman, "Combined low-cost, high-efficient inverter, peak power tracker and regulator for PV applications," *IEEE Trans. Power Electron.*, vol. 6, no. 1, pp. 73–82, Jan. 1991.

[9] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Trans. Industry Appl.*, vol. 48, no. 3, pp. 1064–1069, May/Jun. 2012.

[10] "A regional first: New englanders used less grid electricity midday than while they were sleeping on April 21," 2018. [Online]. Available: http://isonewswire.com/updates/2018/5/3/a-regional-first-new-englanders-used-less-grid-electricity-m.html

[11] Y. Li *et al.*, "Non-detection zone analytics for unintentional islanding in distribution grid integrated with distributed energy resources," *IEEE Trans. Sustain. Energy*, 2018.

[12] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, no. 2., Cambridge, U.K.: Cambridge Univ. Press, 1928, pp. 180–190.

[13] D. Lam, C. Thompson, and J. Croke, "Improving at-site flood frequency analysis with additional spatial information: A probabilistic regional envelope curve approach," *Stochastic Environmental Res. Risk Assessment*, vol. 31, no. 8, pp. 2011–2031, 2017.

[14] D. Maposa *et al.*, "Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalised Pareto distribution," *Statist. Interface*, vol. 10, no. 1, pp. 131–144, 2017.

[15] B. Raggad, "Statistical assessment of changes in extreme maximum temperatures over Saudi Arabia, 1985–2014," *Theor. Appl. Climatol.*, vol. 132, no. 3–4, pp. 1217–1235, 2018.

[16] M. Brunetti, A. Bertolini, M. Soldati, and M. Maugeri, "High-resolution analysis of 1-day extreme precipitation in a wet area centered over eastern Liguria, Italy," *Theor. Appl. Climatol.*, pp. 1–13, 2018. [Online]. Available: https://doi.org/10.1007/s00704-018-2380-1

[17] H. Wu and H. Qian, "Innovative trend analysis of annual and seasonal rainfall and extreme values in Shaanxi, China, since the 1950s," *Int. J. Climatol.*, vol. 37, no. 5, pp. 2582–2592, 2017.

[18] Y. Liu, D. Chen, S. Li, and P. Chan, "Revised power-law model to estimate the vertical variations of extreme wind speeds in China coastal regions," *J. Wind Eng. Ind. Aerodynamics*, vol. 173, pp. 227–240, 2018.

[19] B. Ruszczak and M. Tomaszewski, "Extreme value analysis of wet snow loads on power lines," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 457–462, Jan. 2015.

[20] D. Waliser and B. Guan, "Extreme winds and precipitation during landfall of atmospheric rivers," *Nature Geosci.*, vol. 10, no. 3, pp. 179–183, 2017.

[21] P. Asadi, S. Engelke, and A. C. Davison, "Optimal regionalization of extreme value distributions for flood estimation," *J. Hydrol.*, vol. 556, pp. 182–193, 2018.

[22] C. Yang, A. A. Thatte, and L. Xie, "Multitime-scale data-driven spatio-temporal forecast of photovoltaic generation," *IEEE Trans. Sustain. Energy*, vol. 6, no. 1, pp. 104–112, Jan. 2015.

[23] Q. Li, Z. Wu, and X. Xia, "Estimate and characterize PV power at demand-side hybrid system," *Appl. Energy*, vol. 218, pp. 66–77, 2018.

[24] Y. Hu, W. Lian, Y. Han, S. Dai, and H. Zhu, "A seasonal model using optimized multi-layer neural networks to forecast power output of PV plants," *Energies*, vol. 11, no. 2, p. 326, 2018.

[25] X. G. Agoua, R. Girard, and G. Kariniotakis, "Short-term spatio-temporal forecasting of photovoltaic power production," *IEEE Trans. Sustain. Energy*, vol. 9, no. 2, pp. 538–546, Apr. 2018.

[26] C. Wan, J. Lin, Y. Song, Z. Xu, and G. Yang, "Probabilistic forecasting of photovoltaic generation: An efficient statistical approach," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2471–2472, May 2017.

[27] A. Stein, R. R. Draxler, G. D. Rolph, B. J. Stunder, M. Cohen, and F. Ngan, "NOAAs HYSPLIT atmospheric transport and dispersion modeling system," *Bull. Amer. Meteorol. Soc.*, vol. 96, no. 12, pp. 2059–2077, 2015.

[28] N. G. Korfe and B. A. Colle, "Evaluation of cool-season extratropical cyclones in a multimodel ensemble for Eastern North America and the Western Atlantic Ocean," *Weather Forecasting*, vol. 33, no. 1, pp. 109–127, 2018.

[29] S. Sah, A. Gaur, and M. P. Singh, "Evaluating pattern classification techniques of neural network using k-means clustering algorithm," in *Next-Generation Networks*, Berlin, Germany: Springer, 2018, pp. 563–588.

[30] K. Wang, X. Qi, H. Liu, and J. Song, "Deep belief network based k-means cluster approach for short-term wind power forecasting," *Energy*, vol. 165, pp. 840–852, 2018.

[31] B. Jain, G. Brar, and J. Malhotra, "EKMT-k-means clustering algorithmic solution for low energy consumption for wireless sensor networks based on minimum mean distance from base station," in *Networking Communication and Data Knowledge Engineering*, Berlin, Germany: Springer, 2018, pp. 113–123.

[32] A. Ly, M. Marsman, and E.-J. Wagenmakers, "Analytic posteriors for Pearson's correlation coefficient," *Statistica Neerlandica*, vol. 72, no. 1, pp. 4–13, 2018.

[33] T. Raziei, "A precipitation regionalization and regime for Iran based on multivariate analysis," *Theor. Appl. Climatol.*, vol. 131, no. 3-4, pp. 1429–1448, 2018.

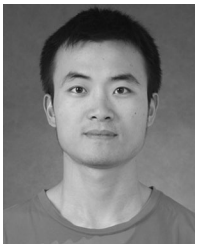
[34] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An Introduction to Statistical Modeling of Extreme Values*, vol. 208. Berlin, Germany: Springer, 2001.

[35] M. Van Montfort and J. Witter, "The generalized Pareto distribution applied to rainfall depths," *Hydrol. Sci. J.*, vol. 31, no. 2, pp. 151–162, 1986.

[36] J. Holmes and W. Moriarty, "Application of the generalized Pareto distribution to extreme value analysis in wind engineering," *J. Wind Eng. Ind. Aerodynamics*, vol. 83, no. 1–3, pp. 1–10, 1999.

[37] C. Y. Huang, C. S. Kuo, and S. P. Luan, "Evaluation and application of bounded generalized Pareto analysis to fault distributions in open source software," *IEEE Trans. Rel.*, vol. 63, no. 1, pp. 309–319, Mar. 2014.

[38] P. Cheng and J. Zhao, "Generalised Pareto distribution-based Bayesian compressed sensing inverse synthetic aperture radar imaging," *IET Radar, Sonar Navigation*, vol. 12, no. 5, pp. 549–556, 2018.
[39] J. R. M. Hosking and J. R. Wallis, *Regional Frequency Analysis: An Approach Based on L-moments*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[40] G. Forzieri *et al.*, "Escalating impacts of climate extremes on critical infrastructures in Europe," *Global Environmental Change*, vol. 48, pp. 97–107, 2018.
[41] H. A. Kazem and M. T. Chaichan, "Effect of humidity on photovoltaic performance based on experimental study," *Int. J. Appl. Eng. Res.*, vol. 10, no. 23, pp. 43 572–43 577, 2015.
[42] M. K. Panjwani and G. B. Narejo, "Effect of humidity on the efficiency of solar cell (photovoltaic)," *Int. J. Eng. Res. Gen. Sci.*, vol. 2, no. 4, pp. 499–503, 2014.
[43] S. Mekhilef, R. Saidur, and M. Kamalisarvestani, "Effect of dust, humidity and air velocity on efficiency of photovoltaic cells," *Renewable Sustain. Energy Rev.*, vol. 16, no. 5, pp. 2920–2925, 2012.
[44] S. Dubey, J. N. Sarvaiya, and B. Seshadri, "Temperature dependent photovoltaic (PV) efficiency and its effect on PV production in the world–a review," *Energy Procedia*, vol. 33, pp. 311–321, 2013.
[45] T. Yang *et al.*, "Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China," *J. Hydrol.*, vol. 380, no. 3–4, pp. 386–405, 2010.
[46] S. A. Khan, I. Hussain, T. Hussain, M. Faisal, Y. S. Muhammad, and A. M. Shoukry, "Regional frequency analysis of extremes precipitation using L-moments and partial L-moments," *Adv. Meteorol.*, 2017, Art. no. 6954902.
[47] J. Fürst, A. Bichler, and F. Konecny, "Regional frequency analysis of extreme groundwater levels," *Groundwater*, vol. 53, no. 3, pp. 414–423, 2015.

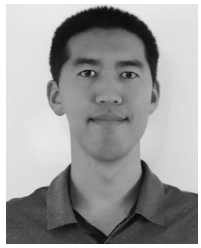


Zefan Tang (S'15) received the B.S. degree in mechanical engineering from Zhejiang University, Zhejiang, China, in 2014, and the M.S. degree in electrical and computer engineering from the University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, China, in 2017. He is currently working toward the Ph.D. degree in electrical engineering with the University of Connecticut, Storrs, CT, USA.

His current research interests include distributed renewable energy systems, cyber physical security for electric power network, machine learning, power system resilience, and cyberattack-resilient load forecasting.

Peng Zhang (M'07−SM'10) received the Ph.D. degree in electrical engineering from the University of British Columbia, Vancouver, BC, Canada, in 2009. He is the Centennial Chair Professor, and an Associate Professor of electrical engineering with the University of Connecticut, Storrs, CT, USA. He was a System Planning Engineer with BC Hydro and Power Authority, Vancouver, BC, USA. His research interests include microgrids, power system stability and control, cyber security, and smart ocean systems.

He is an individual member of CIGRÉ. He is an Editor for the IEEE TRANSACTIONS ON POWER SYSTEMS and the IEEE POWER AND ENERGY SOCIETY LETTERS, an Associate Editor for the IEEE JOURNAL OF OCEANIC ENGINEERING and the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS.



Kunihiro Muto received the B.S. degree in electrical engineering from the University of Connecticut, Storrs, CT, USA, in 2012, where he is currently working toward the M.S. degree in electrical engineering.

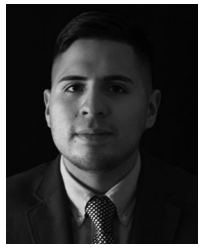
He is a Protection and Controls Engineer with The United Illuminating Company, Orange, CT, USA.

Martial Sawasawa received the B.S. degree in electrical engineering from the University of Connecticut, Storrs, CT, USA, in 2018, and is currently working toward the M.S. degree in electrical engineering with Carnegie Mellon University, Pittsburgh, PA, USA.

He is a Carnegie Mellon University GEM Fellow and a student member of the American Institute of Aeronautics and Astronautics, Reston, VA, USA.

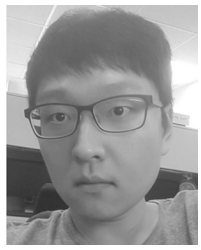
Marissa Simonelli (S'16) received the B.S. degree in electrical engineering with a minor in mathematics from the University of Connecticut, Storrs, CT, USA, in 2018. She is currently working with National Grid, Waltham, MA, USA, in the Protection Engineering Department. Before graduation she interned with ISO New England, Holyoke, MA, USA, in the Transmission Planning Group and with Eversource Energy, Berlin, CT, USA, in the System Operations Group.

She is a student member of IEEEs Power and Energy Society. She was also the 2017–2018 IEEE PES Scholarship Plus Initiative recipient.



Christopher Gutierrez (S'15) received the B.S. degree in electrical engineering from the University of Connecticut, Storrs, CT, USA, in 2018. He is currently working in the Aerospace Industry with Sikorsky, Storrs, CT, USA focusing on manufacturing engineering for various developmental projects for the United States Air Force and United States Marine Corps.

His research interests include power system stability and fabrication of semiconductor devices.



Jaemo Yang received the B.S. degree in civil and environmental engineering with Korea University, Seoul, South Korea, in 2009, the M.S. degree in hydraulics/hydrology engineering with Korea University in 2011, and the Ph.D. degree with the Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA, in 2018. He is currently working toward the Postdoctoral Researcher with the National Renewable Energy Laboratory, Golden, CO, USA in solar forecasting.

His research focuses on developing methodologies that are computationally inexpensive to improve numerical weather prediction of extreme weather events, which are defined by the occurrence of high wind speed and/or intense precipitation (i.e., thunderstorms, train/wind events, tropical storms).

Marina Astitha received the Ph.D. degree in environmental physics from the Physics Department, National and Kapodistrian University of Athens, Athens, Greece, in 2007. She is an Assistant Professor and an Associate Director of the Environmental Engineering program with the Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA. Her research focuses on prediction of extreme weather events, uncertainties and complex error interactions in atmospheric and air quality modeling systems and anthropogenic activities that alter the atmospheric and aquatic environment.

She serves as a Member of the Editorial Board for the Journal of Air & Waste Management Association, a Life Member of the American Geophysical Union, a Life Member of the European Geosciences Union, a Regular Member of the American Meteorological Society, the Association of Environmental Engineers and Science Professors, and the Air and Waste Management Association.

David A. Ferrante received the B.Sc. degree in civil engineering from the University of Vermont, Burlington, VT, USA, and the MBA degree from the University of Hartford, West Hartford, CT, USA. He is the Manager of Distributed Energy Resources and Technology with Eversource Energy, Berlin, CT, USA. Since 2008, he has been leading various corporate and public policy initiatives to integrate distributed energy resources and other advanced smart grid technologies that can integrate and interface with the electric power distribution system.

He and his team have integrated more than 500 MWs of distributed energy resources such as photovoltaic generation, combined heat and power plant, fuel cells, microgrids, battery storage, and electric vehicle charging stations. He worked for more than 15 years in the natural gas industry with Yankee Gas Services Company as a Distribution Engineer, a strategic business account energy consultant, and as the Director of customer service. He is certified by the American Gas Association as an Industrial Consultant and is a Senior Member of the American Association of Energy Engineers and a Registered Professional Engineer in the State of Connecticut.

Joseph N. Debs received the B.Sc. degree in electrical engineering and Master of Business Administration degree from the University of New Haven, West Haven, CT, USA, in 1986 and 1994, respectively. He is the Program Manager of Renewable Resources with Eversource Energy, Berlin, CT, USA.

Robert Manning (M'88) received the B.Sc. degree in electrical engineering from Worcester Polytechnic Institute, Worcester, MA, USA, and the MBA degree from the University of New Haven, New Haven, CT, USA. He is currently a Principal Innovation Engineer with The United Illuminating Company, Orange, CT, USA. He has been with UI since 1988.

He has held various positions in the areas of distribution operations, planning, reliability, and distributed generation. He is a registered Professional Engineer in the State of Connecticut.

James Mader received the B.Sc. degree in electrical engineering technology from Northeastern University, Boston, MA, USA, in 1991. He is currently the Manager of Programs and Projects in the Smart Grids Innovation area with AVANGRID, Orange, CT, USA.

He has worked in multiple engineering disciplines during his 27 years with the United Illuminating including Substation Engineering, Protection & Control Engineering, Transmission Engineering and Distribution design. He is a certified Project Management Professional.