Estimating Power Generation of Invisible Solar Sites Using Publicly Available Data

Hamid Shaker, *Student Member, IEEE*, Hamidreza Zareipour, *Senior Member, IEEE*, and David Wood

***Abstract*—Large-scale integration of invisible solar photovoltaic generation into power systems could significantly affect the system net load and pose new challenges in the operation of power systems. Invisible solar photovoltaic refers mainly to small-scale roof-top solar sites that are not monitored, and thus are invisible to utilities and system operators. Invisible solar generation affects the shape of system net electrical load and could make net load forecasting more challenging. In this paper, a methodology is proposed to estimate the power generation of invisible solar photovoltaic sites. The proposed method only uses the measured power generation data of publicly available sites. It uses real time data of a small subset of sites to estimate the aggregated power generation from known sites within a region. The proposed model is validated using actual invisible solar generation data of the California power system.**

***Index Terms*—Behind-the-meter solar, invisible solar power generation, unsupervised modelling, fuzzy systems.**

I. Introduction

INVISIBLE solar power, mainly in the form of small-scale roof-top photovoltaic (PV) modules, is the capacity that is not monitored by, and thus not visible to, power system operators. Invisible solar PV capacity has grown substantially over the past decade. The worldwide small-scale roof-top solar power installation was 23 GW at the end of 2013, and it is estimated to continue growing at above 20 additional GW per year until 2018 [1]. In the US, California is the pioneer in solar PV capacity with 3,217.0 MW of residential and commercial solar PV installations through 403,504 small-scale projects as of August 2015 [2]. This is about 5% of California's generation capacity.

Significant amount of invisible solar power reshapes the grid's net load pattern [3]. Net load pattern is a key factor in scheduling the short term operation of power systems [4]. Thus, good estimations of invisible solar power generation is necessary to ensure efficient power system operation planning.

Manuscript received October 4, 2015; revised December 15, 2015 and January 22, 2016; accepted January 26, 2016. Date of publication March 4, 2016; date of current version August 19, 2016. This work was supported in part by the Canadian Natural Science and Engineering Research Council, and in part by the ENMAX Corporation under the Industrial Research Chairs Program. Paper no. TSG-01275-2015.

H. Shaker and H. Zareipour are with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB T2N1N4, Canada (e-mail: hshakera@ucalgary.ca; h.zareipour@ucalgary.ca).

D. Wood is with the Department of Mechanical and Manufacturing Engineering, University of Calgary, Calgary, AB T2N1N4, Canada (e-mail: dhwood@ucalgary.ca).

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/TSG.2016.2533164

The literature on estimating invisible solar power generation is very limited. There are reports of attempts by industry-government partnerships to address the issue [5], [6]. However, to the best of the authors' knowledge, no methodology on estimating invisible solar power generation has yet been reported in the literature. In our previous work, we established a basis for invisible solar power estimation [7]. The methodology was based on data reduction techniques and identification of a limited number of solar PV sites whose data could be used to estimate the total power generation from a much larger set of sites. That methodology was developed based on the premise that historical measurements data from every single site over a limited time period was available and used in the model training stage. The developed method in [7] was not designed to adapt to the gradual growth of solar sites. Furthermore, our previous work was not developed to provide any information regarding the uncertainty associated with the estimations. The main source of uncertainty in solar power generation comes from cloud movements, which are difficult to predict.

In the current work, we build on the work in [7] and propose a methodology to estimate power generation from invisible solar PV sites. There are four main distinctions between the model proposed in [7] and the one proposed in the present paper. First, the model in the present work provides information on the uncertainty associated with the estimated power generation volumes. The uncertainties are included by means of fuzzy arithmetic. Second, the model in the present work is unsupervised and hence, does not need the historical measurement data for the training stage. Third, the model proposed in the current paper can easily adopt as the capacity of the invisible solar PV fleet gradually grows. Hence, the model parameters are continuously updated in real time practice. And finally, the current work relies on data from a limited number of sites. That data is publicly available and there is no need for a dedicated data collection system. The model in [7] is a helpful tool in improving the performance of the model proposed in the current paper.

The main contributions of the current paper are as follows. (i) a model is proposed to estimate the invisible solar PV power generation. (ii) the model does not need the historical aggregated generation data during the training stage. (iii) the model only relies on the publicly available data and does not need new measurement devices yet it is powerful and fairly accurate. And (iv), the proposed method can be easily adopted as the capacity in the region grows. The proposed model is validated against actual invisible solar PV generation data from California.

The rest of this paper is organized as follows. First, the proposed methodology in provided Section II. Then, numerical results are presented in Section III. Finally, the paper concludes in Section IV.

II. The Proposed Methodology

In this section, we present the underlying assumptions and methodology for estimating the invisible solar PV power generation.

A. Methodology Overview

It is expected that weather conditions of two close locations are more similar compared to more distant ones [8], [9]. This is the core idea of the proposed approach in this work. For simplicity, temporal relationship of the generation of the sites is not included in the analyses. The reason is that we mainly use neighbouring sites to model other sites. In other words, each city or neighbourhood is modelled separately. This means that time difference has a minor impact on the results. Obviously time difference is very important for high resolution data such as 1s, which is out of the scope of the current study.

Denote the measured output power of site i at time t by $p_i(t)$, for $i \in I$, where I is the set of all roof-top solar PV generation sites whose aggregated generation is to be estimated and let c_i the capacity of site i. In addition, the location of every single site is known and specified by its longitude and latitude, referred to here by Lon_i and Lat_i for site i, respectively. Hence, the distance between each pair of sites i and j is known. We relate $p_i(t)$ and $p_j(t)$ at each time interval t as follows:

$$p_i(t) = \alpha_{ij}(t) \cdot \frac{c_i}{c_j} \cdot p_j(t), \forall i, j \in I, \ \forall t = 1, ..., T \qquad (1)$$

where, $\alpha_{ij}(t)$ in general, is a dimensionless function relating the power generation at the two sites i and j at time t, and T is the total number of modelling time steps. Note that t only includes day time hours since there is no solar PV generation at night. α is defined here to help us identify the relationship between the shape of generation time series of two arbitrary sites.

In a perfect world, α would be a known deterministic function that varies based on time, location and other known factors, such as, site specifications. However, in reality, α is uncertain, and affected by unpredictable cloud movements, changes in composition of the clear atmosphere, variations in inverter performance, module temperature, and soiling. These factors could cause the highest fluctuations in PV power production at time scales less than an hour [8]. But such factors are not easily quantifiable.

There are however two factors that are quantifiable and could be used to estimate α. The first one is the distance between the sites. If the distance of neighbouring sites is sufficiently small, we can estimate their aggregated generation by using the data of at least one representative site. The second factor is the level of power generation at each time. For instance, in early morning the relative variations in generation of one site compared to another could be very large due to small differences in the azimuth angle of the modules. For instance, one site could have 0.01 per unit power output while another close site generates 0.02 per unit at the same time. This means $(0.02-0.01)/0.01 \times 100 = 100\%$ relative power generation difference from the first site's point of view. Hence, we focus on estimating α as a non-parametric uncertain function based on distance of the sites and their generation level.

Define $P_{tot}(t)$ as the measured total power generation for all the sites at time interval t as follows:

$$P_{tot}(t) = \sum_{i \in I} p_i(t), \qquad \forall t = 1, ..., T. \qquad (2)$$

Considering the potential large number of sites in a power system's control area, instead of estimating α for each pair of sites, we break the entire region in smaller subregions where the distances between the site pairs are limited. We cluster the sites of set I into M subregions. Each of these M subregions could represent a city or neighbourhood in a large city. We select one representative site for each of the M subregions. Equation (2) then could be rewritten in terms of the generation level of the M selected sites as follows:

$$P_{tot}(t) = \sum_{m=1}^{M} \left\{ A_m(t) \cdot \frac{p_m(t)}{c_m} \right\}, \qquad \forall t = 1, \ldots, T \qquad (3)$$

where,

$$A_m(t) = \sum_{i \in K_m} \{\alpha_{im}(t) \cdot c_i\}. \qquad (4)$$

K_m is the set of all roof-top PV sites in subregion $m = 1, 2, \ldots, M$ where $\cup_{m=1}^{M} K_m = I$.

Variable α in the above equation is not known for all possible sites i because not all of the sites in set I are visible to the system operator. However, historical measurements of a smaller set $J \in I$ sites for T time steps is publicly available. Hence, the $m = \{1, 2, ..., M\}$ representative sites could be selected among the set of visible sites J such that they represent the highest possible information of all of the sites.

One approach to choose the representative sites for each subregion is to apply the methodologies proposed in [7] to the data of set J. A hybrid k-means+PCA methodology is proposed in [7] and has shown promising ability in selecting the best informative sites among a large set of sites. This approach uses historical generation data of the sites in a subregion and ranks them using Principal Component Analysis (PCA). Then, the highest ranked sites will be selected considering their distance from the centre of the corresponding subregion. For more information the interested reader can refer to [7].

If the set of J is large enough, the statistical behaviour of all the sites, i.e., in set I, could be estimated using that of set J. Thus, we can build frequency distributions of α using the available data from the sites of set J in each of the M subregions.

One well known approach to model uncertainty is probability theory. However, it can be computationally expensive. The reason is that in probabilistic approaches the calculations are repeated for the most likely or even all of the combinations of

different scenarios. Another approach to quantify uncertainties and vagueness is the implementation of fuzzy numbers and tracing the propagation of uncertainties in the systems through fuzzy arithmetic [10]. Fuzzy arithmetic has been used in a number of applications and has shown promising results. For instance, an arithmetical fuzzy approach for DC load flow problem was presented in [11]. Moreover, a fuzzy approach was proposed to model the reliability of Phasor Measurement Units (PMU) in [12]. In addition, a fuzzy model was proposed to model dynamic thermal rating of transmission lines to take into account weather variability and uncertainty in the calculations [13]. It has been complemented by modelling the reliability of dynamic thermal rating using fuzzy numbers in a fuzzy optimization model [14]. Thus, we have chosen fuzzy arithmetic for modelling the uncertainties.

A fuzzy set [15] is a set whose members have a membership degree, usually in the interval of [0, 1] [16]. A fuzzy number is a special case of fuzzy set whose members are within the set of real numbers. The membership degrees for all of the members of a fuzzy number form a membership function. This membership function must have one and only one increasing segment, and one and only one decreasing segment with minimum and maximum membership degrees of 0 and 1, respectively [10]. For more information on fuzzy numbers and fuzzy arithmetic see [10], [16], [17].

Define $\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t)$, $\tilde{\boldsymbol{A}}_{\boldsymbol{m,g}}$, and $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$ as fuzzy numbers and $\left(\sum\right)$ as fuzzy summation operation. Further, g is the index for the generation level of site m. Hence, equations (3)-(4) are converted to fuzzy equations as follows:

$$\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t) = \left(\sum_{\boldsymbol{m=1}}^{\boldsymbol{M}}\right)\left\{\tilde{\boldsymbol{A}}_{\boldsymbol{m,g}} \cdot \frac{p_m(t)}{c_m}\right\}, \quad \forall t = 1, ..., T, \ \forall g. \tag{5}$$

and

$$\tilde{\boldsymbol{A}}_{\boldsymbol{m,g}} = \left\{\sum_{i \in K_m} c_i\right\} \cdot \tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}} = C_m \cdot \tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}} \tag{6}$$

where, C_m is the total invisible solar PV capacity that is represented by site m. Hence, for each time step we will have a fuzzy number that represents the total generation of all the sites at time t. The procedure of calculating $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$ is discussed next.

Note that as long as the real time invisible solar PV capacities C_m in the M subregions are known, they can be directly included in the proposed model by simply updating C_m at any time in equation (6). Hence, the proposed method easily adapts as the capacity in any subregion changes.

B. Developing Fuzzy Number From Frequency Distributions

One can construct a fuzzy number from frequency distributions using different approaches. In most practical cases, however, it is sufficient to simply normalize the envelope of frequency distributions to the maximum of unity [10]. In the current work, we propose the following procedure to develop $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$ using the frequency distributions of α for the corresponding sites within each of the M subregions, as follows:

1) For all $m = \{1, ..., M\}$, calculate $\alpha_{ij}(t)$ for all combinations of sites i and j as follows:
$$\alpha_{ij}(t) = \frac{c_j}{c_i} \cdot \frac{p_i(t)}{p_j(t)}, \quad \forall t = 1, \ldots, T, \ \forall i, j \in \{K_m \cap J\}. \tag{7}$$
2) Initiate generation level increment $\phi_{gen} = 0$ and $g = 1$. ϕ_{gen} represents the interval length for the per unit generation of site j and g is the index for generation level of site j.
3) While $\phi_{gen} < 1$ repeat the following steps.
4) Build the frequency distribution of all α that are associated with the following condition.
$$\frac{p_j(t)}{c_j} \in \left\{(0, \phi_{gen}^{+}] + \phi_{gen}\right\}, \quad \forall j \in \{K_m \cap J\} \tag{8}$$
where, ϕ_{gen}^{+} is the first discrete level of generation at site m in per units. In this work we use $\phi_{gen}^{+} = 0.1$.
5) Normalize the envelope of frequency distributions to the maximum of unity. If the resulting curve satisfies the characteristics of a fuzzy number, use it as $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$. Otherwise, build a fuzzy number that closely resembles the distribution. In the current work, moving average smoothing made the trend of the distribution acceptable to be a fuzzy number in most of the cases. In other cases, however, the distribution could be modified at the discretion of the modeller. The following checks and modifications need to be done on the resulting fuzzy numbers.
   - There must be one and only one maximum membership degree of 1 in $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$, which occurs at point $\alpha = 1$.
   - The membership function of $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$ must be monotonically increasing prior to point (1,1) and monotonically decreasing after that point.
   - After smoothing, membership degrees for α out of the range of the original frequency distribution must be zero.
6) $\phi_{gen} = \phi_{gen} + \phi_{gen}^{+}$ and $g = g + 1$.

Following the above procedure all of the required fuzzy numbers $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$ will be calculated. For example, if ten generation levels are included in the model, $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{3,7}}$ is the fuzzy number associated with site $m = 3$ and for when this site generates within (0.6, 0.7] per unit.

In addition, for the case when the representative site m is generating at its maximum capacity, we need to use a new fuzzy number $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,11}}$ that is the same as $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,10}}$ but has zero membership degrees for $\alpha > 1$. This guarantees the estimated total power generation level $\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t)$ is never higher than the total installed capacity of all the invisible solar PV sites.

C. Using B Additional Representative Sites

Solar PV generation is very sensitive to passing clouds or anything that blocks the sunshine over the PV modules. Hence, the output pattern of a single site looks very noisy.

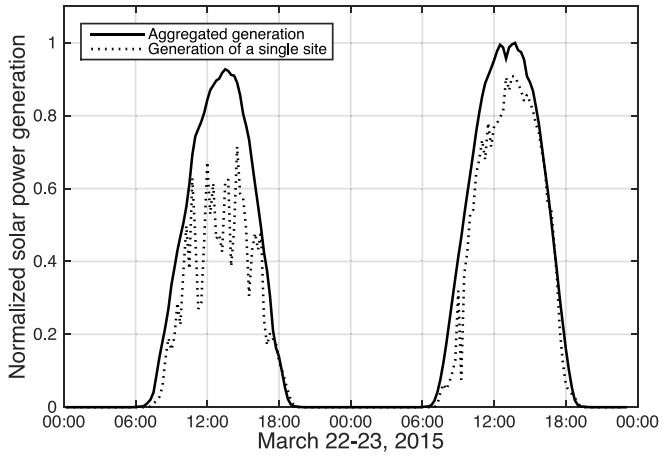

Fig. 1. Aggregated generation vs. a single site's generation of two typical days in California.

However, the aggregated output of a number of sites will have a much smoother shape compared to the individual sites. Figure 1 shows the normalized output of a single site and compares it to the aggregated generation of 6,673 sites in California on two consecutive days. The figure shows that the generation pattern of a single site is very volatile compared to the aggregated values. Since the model is based on a limited number of sites M, the sharp output of those sites will affect the final estimated $\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t)$. Thus, we need to do some smoothing to make the output representative of a larger number of sites.

To remedy this issue it is proposed to use data of B additional sites for every site m and modify the M number of inputs $\frac{p_m(t)}{c_m}$ in equation (5). These $B \times M$ additional sites were chosen within the area of each site m, i.e., K_m based on the highest ranks that the hybrid k-means+PCA methodology [7] on the J sites calculates. Since the proposed model uses only publicly available data, it is easy to have some additional sites included in the fuzzy model with no extra cost. Hence, M number of $\frac{p_m(t)}{c_m}$ inputs to the model in (5), which are the normalized outputs of the M sites will be modified, using the following equation:

$$\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t) = \left(\sum_{\boldsymbol{m=1}}^{\boldsymbol{M}}\right)\left\{\tilde{\boldsymbol{A}}_{\boldsymbol{m,g}} \cdot \left(\frac{p_m(t)}{c_m}\right)^{mdf}\right\}, \quad \forall t = 1, \ldots, T, \quad \forall g, \tag{9}$$

and

$$\left(\frac{p_m(t)}{c_m}\right)^{mdf} = \frac{1}{B+1} \cdot \left(\sum_{b \in B_m}\left\{\frac{p_b(t)}{c_b}\right\} + \frac{p_m(t)}{c_m}\right) \tag{10}$$

where, $B_m \subseteq \{K_m \cap J\}$ is the set of additional included sites in the area of site m. By using (10), the inputs of (9) are smoothed using the additional sites' data. The larger the value of B, the smoother the output would be.

D. Fuzzy Interval of Confidence

In fuzzy arithmetic, the interval of confidence at level λ for a fuzzy number is the interval associated with the fuzzy membership degree of λ. The fuzzy interval of confidence, or simply interval of confidence, shows the truth value of the corresponding parameter for that interval [18]. In that sense we have an interval of confidence for every level of $\lambda \in [0, 1]$. These intervals could be used in applications that consider uncertainty in the modelling procedure.

We use the interval of confidence for the calculated $\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t)$ for all the time instances as follows:

$$\tilde{\boldsymbol{P}}_{tot}^{(\lambda)}(t) = \left[P_{tot_1}^{(\lambda)}, P_{tot_2}^{(\lambda)}\right], \qquad \forall t = 1, \ldots, T, \tag{11}$$

where $P_{tot_1}^{(\lambda)}$ and $P_{tot_2}^{(\lambda)}$ are the lower and upper limits of the total power generation from all the sites at λ level of confidence. As a result, we can have intervals for any desired level of confidence for the estimated total generation from the invisible PV sites.

If the system operator needs only one number as the aggregated invisible solar PV generation, one can use different defuzzification approaches to convert the fuzzy number $\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t)$ to a crisp number. In this work we use *Centroid point* method. For the details of this defuzzification approach the interested reader can refer to [19].

III. Numerical Results

A. Data Description and Bad Data Detection

Most of the roof-top solar power generation technologies provide the user with an interface for optional data sharing. As an example, Enphase [20] is one of the major companies providing micro inverters for residential solar PV systems with such capability. The data used in this work is accessed from the Enphase website that stores 15-minute PV electricity generation data [21]. We have also the approximate location of each site in terms of Longitude and Latitude. All systems in a neighbourhood/city have the same location in the dataset. We collected data from an initial set of 9,777 solar PV sites spread across California. However, after data cleanup and checks, a smaller set of sites was identified which had acceptable data for our analyses.

Four factors could contribute to bad data, namely, measurement equipment failure, loss of internet connection, failure of the local computer, and the PV modules being out of service or unavailable. We have searched the initial raw data set for such issues considering the following criteria:

- Constant output over an extended period of time: A site that reports constant outputs for 2.5 hours for more than 8 days is removed from the initial data set.
- No or zero output: A site that reports no output or zero values for a full day, is removed from the initial data set.
- Unavailable sites: some sites have incomplete data prior to a point in time, i.e., a particular date. One reason is that the installation date of the system occurred after the start of data collection. The other possible reason is that the site does not provide publicly available data for that period or is out of service. Such sites are also removed from the initial data set.

After data cleanup and removal of the bad data sites, 6,673 sites were identified with reliable data. The data covered March 11 to May 10, 2015.

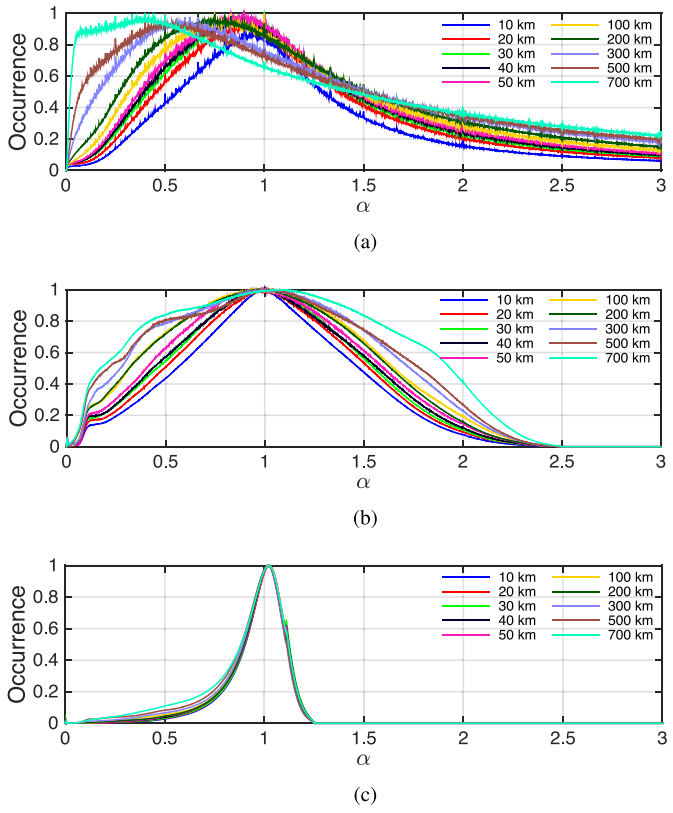

Fig. 2. Frequency distribution of α for different distances: a) Generation decile 1; b) Generation decile 5; c) Generation decile 9.

In Section II-A we distinguished the frequency distributions of α for every distance range and power generation level. To demonstrate how the two parameters impact α, we use the available dataset to generate the frequency distributions of α for all 6,673 sites in California for four of the generation deciles and different distance ranges. The results are shown in Fig. 2. Figures 2(a), 2(b) and 2(c) show the frequency distributions of α for generation deciles 1, 5 and 9, respectively. Here, generation decile refers to the normalized output of site j. For instance, generation decile 5 means that site j is producing 40-50% of its capacity. There is a different behaviour of α in different generation levels, and to some extent, for different distances. Observe that the lowest possible value for any α is zero since power generation cannot go below zero. On the other hand, its maximum is infinity for the case $p_i(t)$ is positive while $p_j(t)$ is zero. This happens during sunrise/sunset hours when not all of the sites start/finish power generation at the same time.

B. Spatial Relationship of PV Solar Generation

To show the importance of generation level, moving from Fig. 2(a) to Fig. 2(c), observe that as the generation level increases, the distributions change from right tailed to left tailed shapes and become much narrower. This shows the importance of including generation level in the modelling procedure. In fact, for different generation levels the general shape of the distributions are totally different.

TABLE I
DETAILS OF THE FOUR CONSIDERED SCENARIOS IN THE CALIFORNIA CASE STUDY; $\eta(.)$ REPRESENTS THE NUMBER OF SITES IN A SET

Scenario name	$\eta(J)$	$\eta(J)/\eta(I)$ [%]
J-840	840	12.6
J-1680	1,680	25.2
J-3360	3,360	50.4
J-6673	6,673	100

Additionally, the other observation from the Fig. 2 is that distance is less important when the level of generation increases, i.e., frequency distributions are more distinguishable in Fig. 2(a) compared to Fig. 2(c). However, in all cases the higher distances make the distributions wider. This is in line with the fact that the closer the sites, the more similar generation pattern they will have. In addition, the peak of the frequency distributions are closer to point $\alpha = 1$ for lower distances. This means that the highest chance is for when the per unit output of site i is the same as the per unit output of site j at the same time. In other words, there is a high chance that if we pick two sites within an area, their per unit generation levels are close to each other. However, for very distinct sites, typically more than 100 km, this may not hold. In fact, from Fig. 2 we can see that for the distances greater than 10 km, peak point of the distributions moves away from $\alpha = 1$. As a result, for the purpose of developing a model using the maximum distance of 10 km is reasonable, which in many cases is grater that the boundaries of cities and neighbourhoods.

C. Converting Frequency Distributions to Fuzzy Numbers

The California data comprises of 42 unique pairs of longitude and latitude with each location containing at least 20 and at most 852 sites. We therefore use $M = 42$ subregions in the simulations. We also use 10 equally distributed generation levels for the fuzzy numbers $\tilde{\boldsymbol{\alpha}}_{m,g}$. Therefore, a total number of $42 \times (10 + 1) = 462$ fuzzy numbers will be developed for the simulations. Moreover, the first 30 days are used for developing the $\tilde{\boldsymbol{\alpha}}_{m,g}$s and the second 30 days are considered as the out of sample test data.

We consider four different scenarios, namely J-840, J-1680, J-3360, and J-6673, each with a different set J with different number of sites that were randomly selected for production of the frequency distributions. Note that all of the scenarios have 42 subregions and the invisible solar capacity is identical among them. The difference between these four scenarios is set J, which indicates the number of considered sites for developing frequency distributions and fuzzy numbers. The number of the sites in each scenario is shown in Table I.

Using the procedure presented in Section II-B, we develop and convert the frequency distributions of α to fuzzy numbers $\tilde{\boldsymbol{\alpha}}_{m,g}$. The fuzzy numbers associated with $\tilde{\boldsymbol{\alpha}}_{\mathbf{6,5}}$ and $\tilde{\boldsymbol{\alpha}}_{\mathbf{33,5}}$ are presented in Fig. 3(a) and Fig. 3(b), respectively, both for Scenario J-840. The button plot of Fig. 3(c) is associated with $\tilde{\boldsymbol{\alpha}}_{\mathbf{33,5}}$ and scenario J-6673 that has higher sample data for the frequency distribution. As Fig. 3 shows, although the

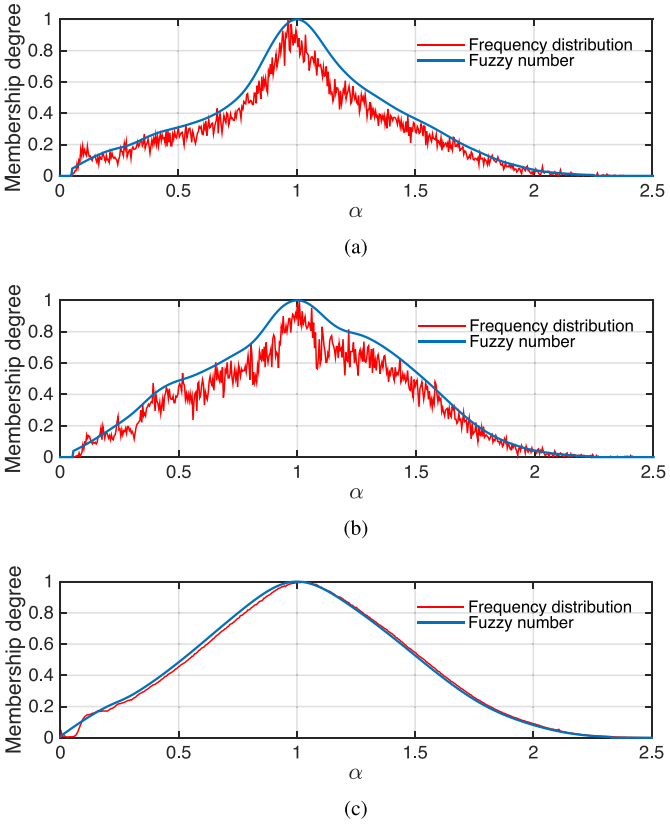

Fig. 3. Three representative frequency distributions of α and $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$: a) Scenario J-840, m = 6, and g = 5; b) Scenario J-840, m = 33, and g = 5; c) Scenario J-6673, m = 33, and g = 5.

frequency distributions look very noisy in some situations, the fuzzy numbers are built such that they are as close as possible to the distributions while satisfying the requirements of a fuzzy number. This proves the efficiency of the proposed procedure in Section II-B.

To show the importance of considering the M subregions, observe from the figure that the shapes of Fig. 3(a) and Fig. 3(b) are different from each other. This means that, as expected, the behaviour of sites are not necessarily the same for two different subregions. Hence, the geographical clustering of the sites is justified. Note that these differences were seen for other m and g values as well but we only showed $\tilde{\boldsymbol{\alpha}}_{\mathbf{6,5}}$ and $\tilde{\boldsymbol{\alpha}}_{\mathbf{33,5}}$ as a sample.

In order to evaluate the impact of the number of sites in building frequency distributions of α we compare Figs. 3(b) and 3(c). The number of sites used to develop $\tilde{\boldsymbol{\alpha}}_{\mathbf{33,5}}$ in scenario J-840 was 20 in Fig. 3(b) while 852 sites were used for $\tilde{\boldsymbol{\alpha}}_{\mathbf{33,5}}$ in Fig. 3(c). As can be seen in Fig. 3(c), having large enough samples for the frequency distribution leads to a much smoother frequency distribution. However, there is not much difference between using 20 and 852 sites in developing $\tilde{\boldsymbol{\alpha}}_{\mathbf{33,5}}$, i.e., the membership functions in Figs. 3(b) and 3(c) are very similar. Hence, using more than about 20 sites may not be beneficial. For the best results though it is suggested to select the sites in different locations of a subregion to have a better representation of potential generation patterns within it.

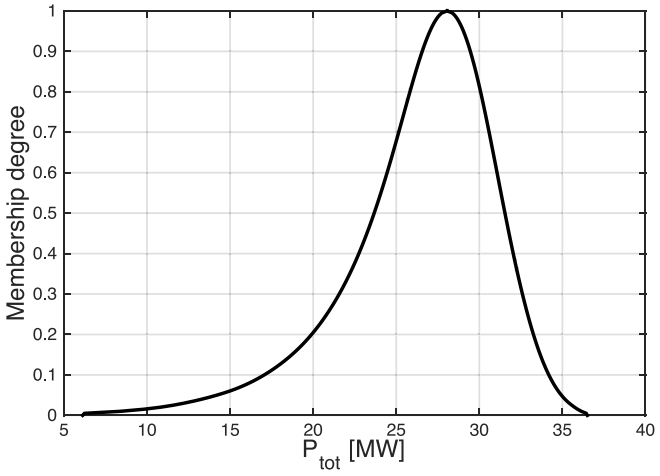

Fig. 4. A sample of a fuzzy number output $\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t)$.

D. Estimation Accuracy

In this section we evaluate the performance of the proposed model in estimating invisible solar power generation.

1) For the Case With Only One Selected Site Per Subregion, i.e., $B = 0$: After generating the required fuzzy numbers we have only used the data from 42 representative sites and calculated the fuzzy aggregated generation from all 6,673 sites. In other words, $B = 0$ at this stage. Note that for the sake of a fair comparison, these selected sites are the same for all of the four scenarios. We applied the hybrid k-means+PCA data reduction approach in [7] to choose the best representative sites for each of the 42 subregions. We also examined a few other ad-hoc methods for such selection (e.g., randomly choosing sites from each region), but the results based on hybrid k-means+PCA data reduction approach were significantly more accurate, and thus, are the only ones presented in the paper. Note that no relevant literature exists on this topic and thus, designing a few ad-hoc methods was the only available solution here.

Figure 4 depicts a sample output $\tilde{\boldsymbol{P}}_{\boldsymbol{tot}}(t)$ of the proposed model for scenario J-840 for one time interval. While the actual $P_{tot}(t)$ for this figure was 28.38 MW, the calculated output for membership degree one was found to be 28.06 MW, which is very close to the actual value. The defuzzified number was 27.62 MW for this particular time step, which shows $(28.06 - 27.62)/27.62 \times 100 = 2.7\%$ error. From Fig. 4, one can obtain the associated interval for each level of confidence. For instance, the total aggregated generation of all the sites is within the interval of $[27.02, 29.04]$ MW with the confidence level of 0.95 and the least level of confidence indicates the most extreme possible output of $[6.12, 36.50]$ MW.

Sample daily estimations with good and weak performances are presented in Figs. 5(a) and (b), respectively. As Fig. 5(a) shows, during sunny days the estimations are very accurate for almost all time. However, on a cloudy day the model may incorrectly estimate the aggregated generation at some times; see Fig. 5(b). In addition, the calculated output is more noisy compared to the actual aggregated values. This comes from the model structure, which is essentially developed based on an educated scaling up of the generation of the M representative

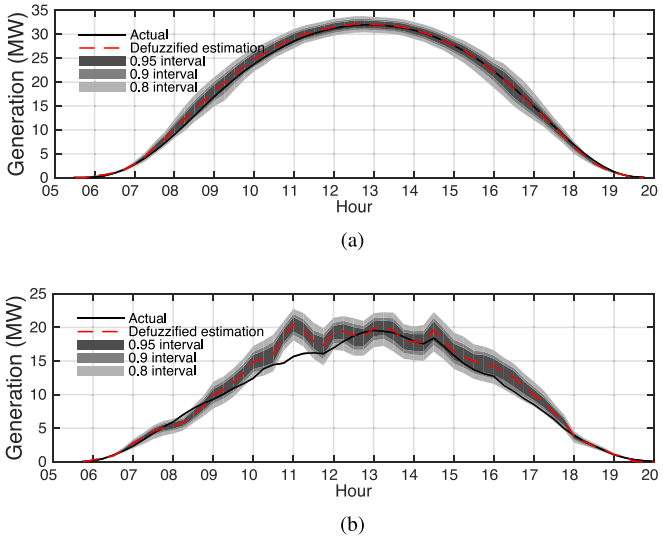

Fig. 5. Defuzzified estimation along with different intervals of confidence of P_{tot} for scenario J-840 and $B = 0$: a) April 27, 2015; b) May 7, 2015.

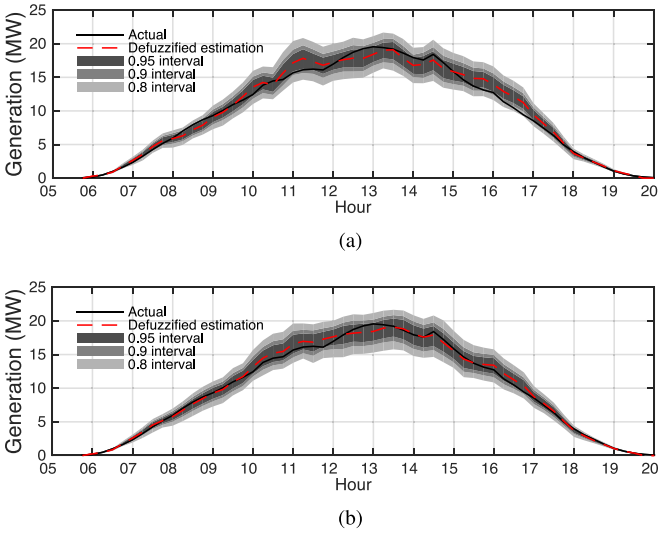

Fig. 6. Defuzzified estimation along with different intervals of confidence of P_{tot} for scenario J-840 on May 7, 2015: a) B = 1; b) B = 3.

sites using $\tilde{\boldsymbol{A}}_{m,g}$. As a result, generation variability of the representative sites will be apparent in the calculated output. In this case, the estimations of cloudy days may be less accurate compared to sunny days. However, note that the model output follows the variations of the actual aggregated generation.

The other observation from Fig. 5 is that the 0.95 interval of confidence includes the actual $P_{tot}(t)$ at all times of the sunny day but misses at some times of the cloudy day. It also shows that the level of confidence is narrower for lower generation levels. This is true since at low generations the absolute uncertainty is less than the intervals with high potential solar energy.

2) For the Case Where More Than One Site is Selected Per Subregion, i.e., B >0: To remedy the issue of estimation errors during cloudy days, different values of $B = 1, ..., 9$ have been included in the model. Figure 6 represents the results for $B = 1$ and 3 on the same day as the one in Fig. 5(b). Observe that using even one additional site for each site m and feeding the modified inputs to the fuzzy model will improve the accuracy of estimations. The issue of high errors in cloudy days is very well resolved.

Comparing Fig. 6(a) with Fig. 5(b) it can be seen that with using only one additional site for each subregion the estimations become much better and the unwanted noise in the estimations distinguishes. Moreover, observe that Fig. 6(a) and (b) show similar results in terms of the general shape of the estimations. This reveals that increasing the number of additional sites over one in each subregion may not improve the results significantly.

For the defuzzified values $\hat{P}_{tot}(t)$, average Daily Root Mean Squared Error (DRMSE) is the measure of accuracy in this work and is defined as follows:

$$DRMSE = \frac{1}{ND}\sum_{d=1}^{ND}\left\{\sqrt{\frac{1}{N_d}\sum_{t\in T_d}\left[P_{tot}(t) - \hat{P}_{tot}(t)\right]^2}\right\}. \quad (12)$$

In equation (12) ND is the total number of test days, and N_d and T_d are the number of daylight hours and the set of all daylight time intervals on day d, respectively.

Table II summarizes the DRMSE of different scenarios and B values. It also compares the best results of model in [7], i.e., linear model (Lin) with the present work. The linear model is a simple yet accurate model that finds a linear function between the realtime generation data of the selected sites and the aggregated generation of all the sites within the considered subregion. In other words, based on this model, the aggregated output of all the sites is a linear weighted average of output data from the selected sites. For more information on this model refer to [7]. The peak generation over the 30 days test period was 32.04 MW. We consider days with less than 25 MW of aggregated daily peak to be cloudy days. Among the 30 test days, nine days are considered cloudy and the rest 21 days are sunny days. As the table shows, for the base case of $B = 0$ and scenario J-840, the DRMSE is 1.7% and 3.2% of the peak generation for sunny and cloudy days, respectively. The overall DRMSE of the base fuzzy model is almost twice as big compared to that of the linear model. However, for $B > 1$ the results of the proposed model are even better than the linear model during cloudy days. Note that the fuzzy model is unsupervised and in general it is expected to have higher errors compared to a supervised model.

Table II also shows that increasing the number of sites in development of fuzzy numbers $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{m,g}}$, i.e., moving from Scenario J-840 to Scenario J-6673, does not lead to any noticeable improvements in the results. For example, in sunny days and for $B = 0$, the DRMSE is 0.55 MW and 0.54 MW for Scenario J-840 and J-6673, respectively. The numbers are highlighted by bold font in the table. This shows a marginal improvement. Hence, using 20 sites for each generation centre is enough to model the fuzzy numbers.

In addition, observe from Table II that for sunny days, the approach may not necessarily benefit from the additional B sites from the DRMSE point of view. However, $B = 4$ is the

TABLE II
DRMSE IN MW, FOR DIFFERENT SCENARIOS AND B VALUES

B	0	1	2	3	4	5	6	7	8	9
Sunny days										
J-840	**0.55**	0.67	0.63	0.57	0.51	0.60	0.66	0.65	0.63	0.69
J-1680	0.54	0.60	0.56	0.50	0.42	0.52	0.58	0.57	0.54	0.59
J-3360	0.54	0.60	0.56	0.50	0.42	0.52	0.58	0.57	0.54	0.59
J-6673	**0.54**	0.59	0.55	0.50	0.42	0.52	0.57	0.56	0.54	0.58
Lin [7]	0.23									
Cloudy days										
J-840	**1.02**	**0.60**	0.48	0.40	0.42	0.43	0.44	0.40	0.41	0.40
J-1680	1.06	0.60	0.46	0.38	0.37	0.38	0.38	0.34	0.35	0.34
J-3360	1.05	0.60	0.46	0.38	0.37	0.38	0.39	0.35	0.35	0.35
J-6673	1.05	0.60	0.46	0.38	0.37	0.38	0.38	0.35	0.35	0.34
Lin [7]	0.51									

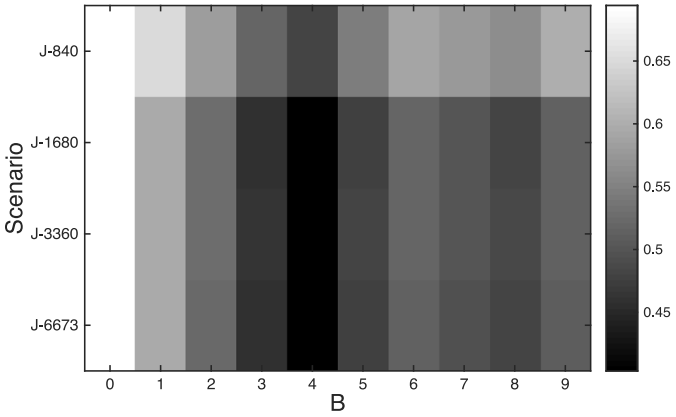

Fig. 7. Heat map of DRMSE in MW, for different scenarios and B values.

TABLE III
PERCENTAGE OF MISSES FOR THE INTERVAL OF CONFIDENCE 0.95, FOR DIFFERENT SCENARIOS AND B VALUES

B	0	1	2	3	4	5	6	7	8	9
Sunny days										
J-840	**11.1**	**7.4**	6.3	7.6	3.3	8.7	11.1	13.0	5.0	3.4
J-1680	11.6	7.2	6.0	7.4	3.0	8.7	11.6	12.5	4.5	3.5
J-3360	11.4	7.2	6.1	8.0	3.2	9.1	12.1	12.9	5.2	4.1
J-6673	11.6	7.3	6.1	7.6	3.0	8.9	11.7	12.8	4.9	3.8
Cloudy days										
J-840	**43.9**	**18.1**	7.1	3.7	2.5	3.4	2.5	1.7	0.3	1.1
J-1680	46.5	19.8	8.2	5.7	2.8	3.4	2.3	1.7	0.6	0.8
J-3360	46.5	21.0	8.5	5.9	3.1	3.7	2.3	2.0	0.6	1.1
J-6673	46.2	20.4	8.2	5.9	2.8	3.7	2.3	1.4	0.6	1.1

optimal choice for them. On the other hand, in general, the higher the B, the better the results would be for the cloudy days up to $B = 4$ and there is not much of improvement for higher B values. Inclusion of at least one additional site for modification of the input from each site m could decrease the DRMSE of cloudy days from 1.02 MW to 0.60 MW, i.e., 41% less. These DRMSE values are shown by bold font in Table II. Figure 7 depicts the heat map of DRMSE for all test days. From this figure it is obvious that $B = 4$ is the optimal case from the DRMSE point of view.

In order to analyze the results in more detail, the percentage of misses for interval of confidence 0.95 and generation levels of higher than 5 MW are summarized in Table III. Observe from the table that the increase of the sites in set J does not necessarily show an improvement in the results. Hence, the

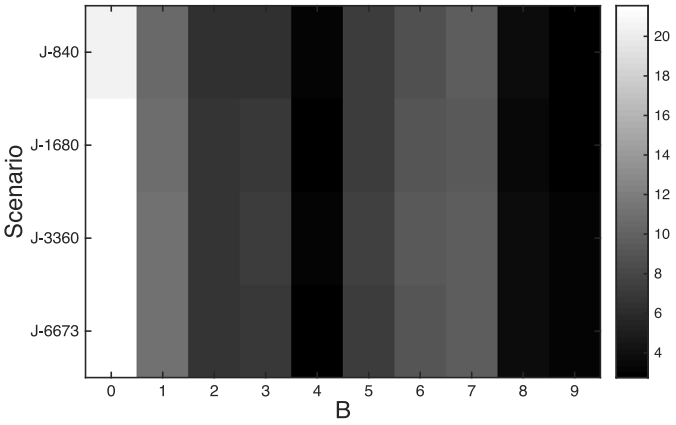

Fig. 8. Heat map of misses for the interval of confidence 0.95, for different scenarios and B values.

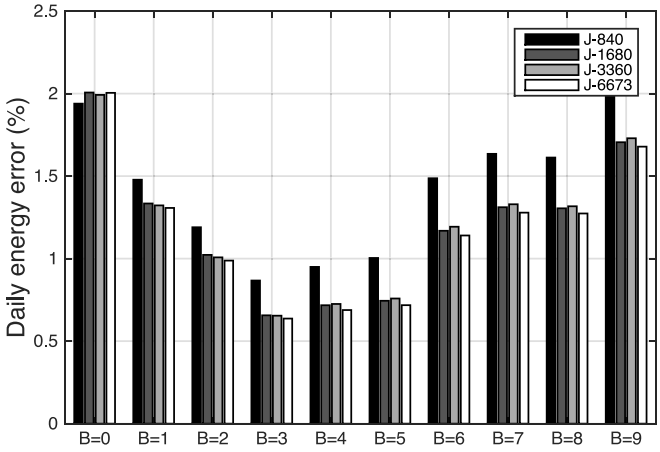

Fig. 9. Daily energy estimation errors of the defuzzified outputs for different scenarios and B values.

assumption of using at least 20 sites for each subregion is justified here as well. However, increasing the B will improve the results significantly for cloudy days compared to sunny days. For instance, in Scenario J-840 when comparing the case of $B = 0$ with $B = 1$, the percentage of misses decreases from 11.1 to 7.4 for sunny days but decreases from 43.9 to 18.1 for cloudy days. This is due to the fact that if at any given time the generation of one of the selected M sites is different from most of the other sites in their subregion, which could be due to passing of cloud over them, the final results would not suffer. The reason is that the generation from the other B sites compensate the unusual output of the site m. However, as the table shows, in all sunny days and some cloudy, the error increases for $B > 4$. The reason behind this fact is that the model is primarily developed based on the generation of the M representative sites. Once we modify the inputs using B additional sites, the model moves to the case that the inputs are no longer the same as those the model was tuned for. Figure 8 also complements the previous discussion by showing the heat map of misses for all of the test days. From the figure one can observe that $B = 4$ is the optimal choice. Although results of $B = 8$ and 9 show a good performance, they show no benefit compared to $B = 4$ in this regard.

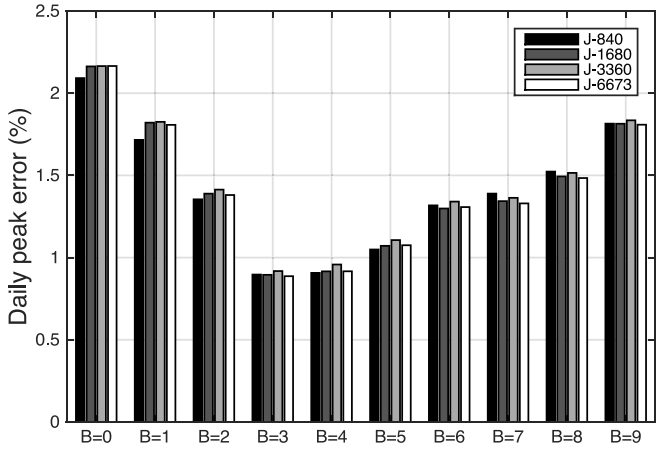

Fig. 10. Daily peak estimation errors of the defuzzified outputs for different scenarios and B values.

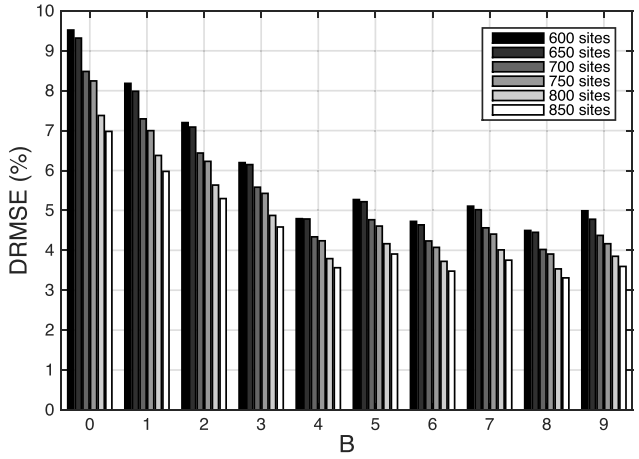

Fig. 11. DRMSE for subregion 33 with different number of installed sites and B values for Scenario J-840.

In addition, the daily energy and peak estimation errors of the defuzzified outputs are depicted in Figs. 9 and 10, respectively. As both of the figures show, there is a constant improvement in the accuracy of estimation as we increase the B up to 3. However, here also the results confirm that high B is not a proper choice for the model. Hence, we can conclude that the use of $B = 3$ or 4 would be the optimal choice. In other words, in order to accurately estimate the power generation of invisible solar PV sites in a city/neighbourhood, real time data of 4 to 5 sites would be enough. In addition, observe from the figures that the difference between different scenarios with equal B is marginal, which is in line with the previous figures and tables and confirms the adequacy of using 20 sites to develop $\tilde{\boldsymbol{\alpha}}_{m,g}$.

E. Effect of Capacity Growth on the Results

To demonstrate that the proposed method can be adjusted as the capacity in any subregion grows, we have chosen subregion 33, which is the largest subregion in our dataset. We randomly selected different number of sites from 600 to 850 within the subregion and estimated the aggregated generation for this subregion. For the sake of a fair comparison, we have used the same selected sites in all cases. Moreover, all of the results are associated with Scenario J-840 which means only 20 sites were used to build the fuzzy numbers. Figure 11 presents the normalized DRMSE for different cases. The DRMSEs are normalized based on the installed capacity of each case. As the figure shows, not only the proposed approach is capable of performing estimations in different capacities, increasing the capacity of the subregion improves the estimation accuracy. The reason for this improvement is the increased smoothing and less unusual pattern in the aggregated generation for higher number of sites. In addition, in line with the previous results, the figure also shows that $B = 4$ is the optimal case in different installed capacities.

IV. Conclusion

In this paper, a methodology is proposed to estimate the power generation of invisible solar photovoltaic sites using a small set of selected representative sites. In the first stage, a subset of sites is selected and within the specified subregions, the possible normalized variations on the generation pattern of one site to another at the same time is calculated. This information is then used to develop fuzzy numbers associated with each subregion. Then, a very small number of representative sites are chosen and their real time power generation is fed to the model as the inputs. Finally, the model calculates a fuzzy number associated with the real time power generation from all the sites within the region based on the existing invisible solar PV capacity of each subregion. To improve the performance of the model, real time generation of additional sites have also been used to modify the inputs of the proposed model.

We have used publicly available data from the Enphase website for the state of California. The results showed that the proposed approach is capable of estimating the generation of invisible sites very well both during sunny and cloudy days. In addition, with the use of data from four additional sites for each subregion, the accuracy of the outputs improves significantly. The results also showed that the use of 20 sites for each subregion would be enough for accurate model parameter estimations. As a result, the fuzzy model could be tuned up for any region with even very large number of invisible solar sites without needing to use all of the sites' data for the modelling procedure. This is a very important practical advantage of the proposed approach.

Acknowledgment

The authors would like to thank Mr. Connor Scheu for his help in preparing the data for the work. Furthermore, We would like to thank the CloudScrape team, especially Mr. Henrik Hofmeister, for his valuable support of this project.

References

[1] *Global Market Outlook for Photovoltaics 2014-2018, EPIA*. [Online]. Available: http://www.cleanenergybusinesscouncil.com/site/resources/files/reports/EPIA_Global_Market_Outlook_for_Photovoltaics_2014-2018_-_Medium_Res.pdf, accessed Nov. 1, 2015.

[2] *Go Solar California*. [Online]. Available: http://gosolarcalifornia.org, accessed Aug. 28, 2015.

[3] K. Chaiamarit and S. Nuchprayoon, "Impact assessment of renewable generation on electricity demand characteristics," *Renew. Sustain. Energy Rev.*, vol. 39, pp. 995–1004, Nov. 2014.
[4] M. Huber, D. Dimkova, and T. Hamacher, "Integration of wind and solar power in Europe: Assessment of flexibility requirements," *Energy*, vol. 69, pp. 236–246, May 2014.
[5] J. S. John. (Oct. 2013). *Transforming Rooftop Solar From Invisible Threat to Predictable Resource*. [Online]. Available: http://www.greentechmedia.com/articles/read/Turning-Rooftop-Solar-from-Invisible-Threat-to-Predictable-Resource.
[6] Innovations-Solar-Energy. *Solar Utility Networks: Replicable Innovations in Solar Energy*. [Online]. Available: http://energy.gov/eere/sunshot/solar-utility-networks-replicable-innovations-solar-energy, accessed Apr. 9, 2015.
[7] H. Shaker, H. Zareipour, and D. Wood, "A data-driven approach for estimating the power generation of invisible solar sites," *IEEE Trans. Smart Grid*, to be published.
[8] M. Lave, J. Kleissl, and J. S. Stein, "A wavelet-based variability model (WVM) for solar PV power plants," *IEEE Trans. Sustain. Energy*, vol. 4, no. 2, pp. 501–509, Apr. 2013.
[9] M. B. P. de Camargoa and K. G. Hubbard, "Spatial and temporal variability of daily weather variables in sub-humid and semi-arid areas of the United States high plains," *Agric. Forest Meteorol.*, vol. 93, no. 2, pp. 141–148, Aug. 1998.
[10] M. Hanss, *Applied Fuzzy Arithmetic: An Introduction With Engineering Applications*. Berlin, Germany: Springer-Verlag, 2005.
[11] M. Cortes-Carmona, R. Palma-Behnke, and G. Jimenez-Estevez, "Fuzzy arithmetic for the DC load flow," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 206–214, Feb. 2010.
[12] F. Aminifar, S. Bagheri-Shouraki, M. Fotuhi-Firuzabad, and M. Shahidehpour, "Reliability modeling of PMUs using fuzzy sets," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2384–2391, Oct. 2010.
[13] H. Shaker, M. Fotuhi-Firuzabad, and F. Aminifar, "Fuzzy dynamic thermal rating of transmission lines," *IEEE Trans. Power Del.*, vol. 27, no. 4, pp. 1885–1892, Oct. 2012.
[14] H. Shaker, H. Zareipour, and M. Fotuhi-Firuzabad, "Reliability modeling of dynamic thermal rating," *IEEE Trans. Power Del.*, vol. 28, no. 3, pp. 1600–1609, Jul. 2013.
[15] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.
[16] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic-Theory and Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, 1995.
[17] A. Kaufmann and M. M. Gupta, *Introduction to Fuzzy Arithmetic Theory and Application*. New York, NY, USA: Van Nostrand Reinhold, 1991.
[18] R. A. McCain, "Fuzzy confidence intervals," *Fuzzy Sets Syst.*, vol. 10, nos. 1–3, pp. 281–290, Jan. 1983.
[19] M. Ma, A. Kandel, and M. Friedman, "A new approach for defuzzification," *Fuzzy Sets Syst.*, vol. 111, no. 3, pp. 351–356, May 2000.
[20] *Enphase*. [Online]. Available: https://enphase.com, accessed Aug. 13, 2015.
[21] *Enlighten*. [Online]. Available: https://enlighten.enphaseenergy.com/public_systems, accessed May 12, 2015.

Hamid Shaker (S'12) received the B.Sc. degree from the Isfahan University of Technology, Isfahan, Iran, in 2007; the M.Sc. degree from the Sharif University of Technology, Tehran, Iran, in 2009; and the Ph.D. degree from the University of Calgary, Alberta, Canada, in 2016, all in electrical engineering. His research interests include invisible solar power generation modeling, renewable energies forecasting and integration into power systems, net load analysis, and fuzzy-based modeling.

Hamidreza Zareipour (S'03–M'08–SM'09) received the B.Sc. degree from the K. N. Toosi University of Technology, Tehran, Iran, in 1995; the M.Sc. degree from Tabriz University, Tabriz, Iran, in 1997; and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2006, all in electrical engineering. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada. His current research interests include economics, planning, and management of intelligent electric energy systems in a competitive electricity market environment.

David Wood received the Bachelor's degree in mechanical engineering and the Master's degree in engineering science from Sydney University in 1974 and 1976, respectively, and the Ph.D. degree in aerodynamics from Imperial College London in 1980. He has been a Professor and NSERC/ENMAX Renewable Energy Chair with the Department of Mechanical and Manufacturing Engineering, University of Calgary, since 2010. His research interests are in small wind turbines and other forms of renewable energy including resource assessment and forecasting.