A Data-Driven Approach for Estimating the Power Generation of Invisible Solar Sites

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

***Abstract*—Roof-top solar photovoltaic systems are normally invisible to system operators, meaning that their generated power is not monitored. If a significant number of systems are installed, invisible solar power could significantly alter the net load in power systems. In this paper, a data-driven methodology is proposed to estimate the power generation of invisible solar power sites by using the measured values from a small number of representative sites. The proposed methodology is composed of a data dimension reduction engine and a mapping function. A number of established methods for reducing the dimension of large-scale data is investigated, and a hybrid method based on *k*-means clustering and principal component analysis is proposed. The output of this block provides a small subset of sites whose measured data are used in the mapping function. We have implemented several mapping functions to estimate the total generation power of all sites based on the measured output of the selected subset of sites. Numerical results based on data from California's power system are presented.**

***Index Terms*—Big data, behind-the-meter solar, clustering, data dimension reduction, invisible solar power generation, principal component analysis.**

I. Introduction

INVISIBLE solar power generation refers to the solar power capacity that is not monitored by, and thus not visible to, power system operators. Mainly in the form of small-scale, roof-top photovoltaic (PV) modules, invisible solar power generation is a growing reality in power systems. California Energy Commission and the California Public Utilities Commission, through the Go Solar California campaign, have announced a goal of 3,000.0 MW of solar power installations on homes and businesses by the end of year 2016. The state seems to be well on track to reach this goal; the installed capacity reached 2,374.0 MW of residential and commercial solar PV installations through 246,266 small scale projects at the end of April 2015 [1]. New worldwide small-scale roof-top solar power installations reached 23 GW in 2013, and is expected to continue growing at a rate of above 20 additional GW per year until 2018 [2].

Several factors contribute to the growing trend in adopting solar power in residential and commercial settings. The cost of solar power generation has continuously declined over the past decade. For example, the turnkey cost of solar power systems in California has dropped from 11 $/W in 2007 to 5.5 $/W in 2014 [1]. In some regions, government policy has promoted and in some cases subsidized solar power generation to respond to environmental concerns around electricity generation. For example, in Ontario, Canada, the Government put in place a generous initial $0.8/kWh feed-in-tariff in year 2009 [3] that led to a boom in solar power industry in this province. Technology breakthroughs on the energy storage side of solar generation systems are also expected to be a defining factor in future growth of solar power. The latest announcement by Tesla Motors [4] promises to make its Powerwall [5] accessible and affordable in near future, which will likely make roof-top solar power generation an even more attractive option.

Despite the environmental benefits of solar power generation, rapid growth of invisible solar generation could potentially pose new challenges in power systems operation. Solar power basically adds negative demand to the system, i.e., when it is available, it reduces the system's total electricity demand. At a high penetration rate, solar power generation can thus impact the system's net load pattern significantly. Net load is defined as the conventional electricity load minus the non-dispatchable generation. On the other hand, system net load is a key input when scheduling the short term operation of power systems. Thus, estimating the system net load in presence of significant invisible solar power generation is of interest. In theory, one may estimate the total solar power generation in the system by monitoring and measuring every single solar generation site. However, with the growing number of small solar sites, this may not be practically possible. Building an infrastructure to monitor, collect, archive and manage solar power generation data for every small-scale site could become very costly. On the other hand, concerns around privacy make it difficult for power utilities to make a plausible case for continuously collecting and using solar power data from private sites. Hence, estimating invisible solar power generation could be a challenge for power utilities.

The industry has already started exploring options to bring a level of visibility to roof-top solar power generation [6]. In late 2013, the U.S. Department of Energy approved funding for several projects to deal with this issue. Defined by

Manuscript received June 3, 2015; revised August 25, 2015 and October 25, 2015; accepted November 1, 2015. Date of publication December 4, 2015; 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-00612-2015.

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.2015.2502140

several industry partners, these projects are meant to develop methods for making roof-top solar power generation more predictable [7]. However, to the best of the authors' knowledge, no literature yet exists that addresses the problem of estimating invisible solar power generation in power systems.

In this paper, we propose a methodology to estimate electricity generation from invisible solar PV resources using data mining tools. The idea is to identify 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. The proposed methodology has two main components, i.e., data dimension reduction and output estimation. For data dimension reduction, we use the Principal Component Analysis (PCA) [8], the *k*-means Clustering [9], a hybrid *k*-means+PCA approach proposed in this work, Relief [10], and the Correlation-based Feature Selection (CFS) [11] techniques. For output estimation, we employ a variety of well established methodologies, including, linear regression [12], Kalman filter [13], Multi Layer Perceptrons (MLP) [14], and Wavelet Neural Networks (WNN) [15]. The main contribution of this paper is to investigate how data-mining could be employed to estimate the aggregated solar power generation from a large set of solar power sites without continuously measuring the output of every single site.

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

II. Proposed Methodology

In this section the proposed framework for estimation of invisible solar generation is introduced. We assume that actual, measured data over a limited period of time is available for the large set of PV sites for which the aggregated output needs to be estimated on an ongoing basis. For example, the data could be arranged to be collected for a 4-month period for all sites of interest. This set of limited data is used to identify a small number of informative sites. We demonstrate that those few sites are the only ones that need to be continuously monitored, and they provide enough information to estimate the total generated power from all the sites.

Although the methodologies in this paper are developed assuming there is a clean set of measurements available for the region, this may be a challenge in practice. However, the data that we have used is sufficient to establish a base for invisible solar power estimations and the challenges involved. The proposed methodology will be the basis for the authors' future works that are focused on particularly two issues, i.e., how to minimize the impacts of data availability limitations on model's performance, and how to make the model adaptable to new addition of solar sites in the region. Note that the new generation of solar power production systems give the owners the option to monitor and publish their power production data to public sites. An example is the system provided by Enphase [16]. As the popularity of such systems increases, more and more data will be available in the future, that would certainly help improving invisible solar power estimates.

A. General Framework

The proposed methodology is developed in four stages, as follows:

1) Data Collection From the Entire Solar Power Generation Fleet: In this stage, power generation data and any other information are collected from the available sites within the territory of interest over a limited time period. Our analyses showed that three to four months worth of data was sufficient to provide reasonable results. While collecting such data on a continuous basis may be very costly and impractical, doing so for a limited period of time could be managed. Denote the measured output power of site i at time t by $p_i(t)$, for $i \in \{1, 2, 3, \ldots, I\}$. I denotes the number of solar PV generation sites for which the total output generation is to be estimated. We also assume that 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. In addition to power generation and location information, there could be M other variables for which the measurements are collected (e.g., temperature, cloud coverage, etc.). We refer to those variables by $x_i^{(m)}$, $m \in \{1, 2, 3, \ldots, M\}$. Assume measurements are available for all variables for T time steps. Thus, the complete set of available information for all sites can be defined as follows:

$$\mathbf{S} = \Big\{ p_i(t), Lon_i, Lat_i, x_i^{(m)}(t) | \quad \forall i = 1, \ldots, I, \ \forall m = 1, \ldots, M, \ \forall t = 1, \ldots, T \Big\}. \quad (1)$$

We also define the measured total power generation for all the sites as follows:

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

2) Data Dimension Reduction: In this stage, a data dimension reduction methodology is used to reduce the complete information set $\mathbf{S}$ into set $\mathbf{S}'$, as follows:

$$\mathbf{S}' = \Big\{ p_j(t), Lon_j, Lat_j, x_j^{(m)}(t) | \quad \forall j = 1, \ldots, J, \ \forall m = 1, \ldots, M, \ \forall t = 1, \ldots, T \Big\} \quad (3)$$

where, J is a smaller set of sites whose information could be used to reasonably estimate the total generated power by all I sites. In general, I is very large, whereas $J \ll I$. Observe that depending on the characteristics of the data dimension reduction technique, availability of data, quality of each of the measured variables, and the modeler's other preferences, a subset of the available information, say $\mathbf{S}^{(.)} \subseteq \mathbf{S}$, may be used at this stage.

We have used five selected data dimension reduction methodologies in this work, as explained in Section II-B.

3) Building the Mapping Function: In this stage, the parameters of a mapping function are estimated. The function maps the total power generation from all I sites to the information from the J selected sites. Define the estimated total power generation from all sites by $\hat{P}_{tot}(t)$. The model, referred to here by f, can be generically described as follows:

$$\hat{P}_{tot}(t) = f\Big(\mathbf{S}'\big|_{t,t-1,\ldots,t-lag}\Big) + \epsilon \quad (4)$$

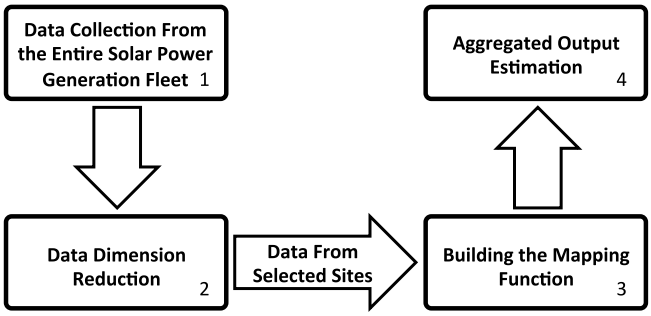

Fig. 1. The proposed framework for invisible solar generation estimation.

where, ϵ is the mapping error. Data of prior times to t may also be included in the inputs. In (4) *lag* indicates the lag times in the input data. The objective of building the model is to estimate the mapping function f such that the difference between the values of $P_{tot}(t)$ and $\hat{P}_{tot}(t)$ are minimized. In this paper, four mapping functions are explored, as explained in Section II-C.

4) Aggregated Output Estimation: In the last stage, measurements from only the J selected sites are used as the input to the mapping function of the previous stage. The output is an estimate of the total power from all I sites. Measurements from other sites are no longer needed. Figure 1 summarizes the above four stages.

B. Employed Data Dimension Reduction Techniques

In this work, we implement five data dimension reduction techniques, namely, Principal Component Analysis (PCA), k-means Clustering, a proposed hybrid k-means+PCA approach, Relief, and Correlation-based Feature Selection (CFS). These methods have been used previously in the literature in similar applications and good performance has been reported [17]–[24]. However, the proposed methodology does not rely on a specific technique.

1) Principal Component Analysis (PCA): PCA is a well-known linear dimension reduction technique. It is very fast and computationally efficient [17]. It constructs a low-dimensional representation of the data that describes the most probable variance in the data [25], [26]. PCA is found to perform better than non-linear complex classification approaches in natural databases [25]. It has been applied in Synchrophasor data dimensionality reduction [17], extracting the distinct features of fault component [18], wind farm a, and grouping of wind parks [19]. More details on formulation and implementation of this method can be found in [8].

The primary output of the PCA method is lower dimensional data. If fact, it uses all of the available information and maps them to the new space using different principal vectors, where each has information amount relative to its corresponding eigenvalue. However, the ultimate goal here is to rank and select the candidate sites. In order to do so, different methodologies could be used. In this work, backward elimination technique is applied [8]. This procedure ranks all the sites based on the information they contain. The reasoning behind backward elimination method is that an Eigen vector related to a small Eigen value correspond to near-constant relationships among a subset of sites. If one of the variables involved in such a relationship is removed, little information will be lost. An obvious choice would be the variable with the highest coefficient in absolute value in the relevant Eigen vector [8].

In the present research work, we only had access to power generation and locational information of each site. No other measured variables were available. We applied the PCA technique to power generation values, i.e.,

$$\mathbf{S}^{\mathrm{PCA}} = \{p_i(t)|\ \forall i = 1, \ldots, I,\ \forall t = 1, \ldots, T\}. \quad (5)$$

2) k-Means Clustering: This technique basically reduces the dimension of the input data by grouping the sites into clusters. The objective is to find clusters that contain as many similar sites as possible yet are different from other clusters [9]. In practice, the number of individual roof-top sites could be very large. Adding the time dimension to the problem, one can appreciate that the number of variables used to find a small number of clusters could easily grow very large, and make the method computationally impractical [27]. Thus, in this work, we have only included the locational information of the sites, i.e.,

$$\mathbf{S}^{\mathrm{KM}} = \{Lon_i, Lat_i|\ \forall i = 1, \ldots, I\}. \quad (6)$$

The rationale is that the sites that are geographically close, are perhaps subject to similar solar regimes and thus, their outputs are close. The idea here is to group the I sites into a small number of cluster of sites, say J, and choose a representative site from each cluster. Applying this technique leads to J clusters. The method provides a centroid for each cluster, which is not necessarily coincident with any of the sites. We choose the site that has the minimum distance to the cluster centroid as the site representing the cluster.

3) Proposed Hybrid k-Means+PCA Approach: Using each of the two previously discussed techniques may not always yield the best outcome. Although PCA is fast and efficient, it sometimes does not find good solutions [25]. Furthermore, since k-means in the current work only relies on the locational data, it only selects the site that is closest to the centroids regardless of the quality of its data. To mitigate these issues, we propose a hybrid k-means+PCA approach to benefit from the strengths of both of these approaches. Hence, the input information set for this approach is as follows:

$$\mathbf{S}^{\mathrm{PA}} = \{p_i(t), Lon_i, Lat_i|\ \forall i = 1, \ldots, I,\ \forall t = 1, \ldots, T\}. \quad (7)$$

The proposed hybrid k-means+PCA technique has the following steps:

Step 1) Apply the k-means Clustering on the I sites and find the J clusters.
Step 2) For each cluster repeat the following steps.
Step 3) Find the Euclidian distance of each site in the cluster to the centroid.
Step 4) Sort the sites based on the calculated Euclidian distance. Then, calculate the cumulative distance of the sites, and normalize the cumulative distances based on the last cumulative distance.

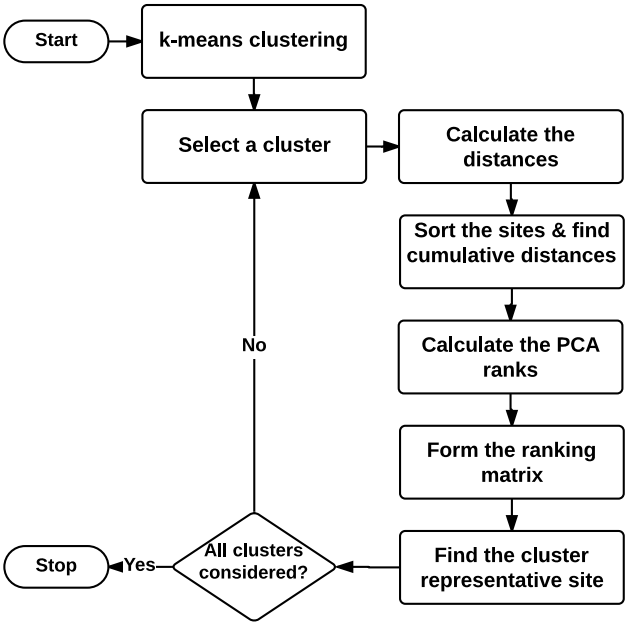

Fig. 2. The proposed hybrid *k*-means+PCA approach.

Step 5) Rank all of the sites in the cluster using the PCA technique.

Step 6) Form a matrix whose first column shows the site numbers in the order from highest to lowest PCA ranks. The second column contains the associated normalized cumulative distances.

Step 7) Choose the first site for which the cumulative distance is less than a pre-specified threshold α, however, the cumulative distance for the next site in the column is more than α. The site associated with the selected order is the representative for the corresponding cluster. α is a measure that forces the selected site to be closer to the centroid compared to at least $(1-\alpha)\times 100\%$ of the sites in the cluster. If such cumulative distance is not found, choose the first site that has lower cumulative distance compared to the next site in the column. If none of the conditions hold, choose the site with the highest PCA rank in Step 5.

This algorithm is represented in Fig. 2. The above steps find the site that takes into account the information from both *k*-means Clustering and PCA methods. Although the priority is based on within-cluster PCA ranks, the threshold α considers the distances such that the site with highest possible PCA rank and closest to the centroid point is chosen.

4) Relief: The Relief is well-established feature selection approach, proposed by Kira *et. al*, and inspired by instance-based learning [10]. In our particular application, it selects the sites that are statistically relevant to the target variable, e.g., $P_{tot}(t)$. By defining an instance from a selected subset of sites, Relief assigns weights for the different sites which are iteratively computed for each randomly selected instance together with its corresponding near-hit and near-miss instances. The closest instance within the same class set is the near-hit instance while the near-miss instance is the closest instance in the opposite class set. It uses Euclidean distance for selecting near-hit and near-miss instances. Finally, relevant sites are chosen when they have weights larger than the specified threshold [10], [24]. Relief was later extended for multi-class classification problems in [28]. Relief was recently implemented in applications such as electricity price forecasting [20], load forecasting [21] and probabilistic modeling [22]. Hence, it was explored as a candidate data dimension reduction technique. We use the power generation at individual sites as the input for Relief, similar to the input for PCA method defined in (5).

5) Correlation-Based Feature Selection (CFS): The idea here is to select the sites that have a high correlation with the target yet having low correlation with each other. This is the basis of the CFS that was introduced by Hall [11]. Because of its good speed and efficiency, it has been used in feature selection applications such as electricity load forecasting [21], [23], and electricity price classification [24]. Hence, we also employ CFS as one of the data dimension reduction techniques. We use generation data as the input to the CFS.

While the data reduction methods may have some similarities, we use them to diversify the numerical experiments and determine how exactly each method may perform in this particular application.

The first three data dimension reduction tasks are performed in MATLAB®. Relief and CFS are implemented using the java-based Waikato Environment for Knowledge Analysis (WEKA) software package [29].

C. Model Estimation

The goal of this stage is to estimate the aggregated generation of all of the I sites, using only the data of J selected sites. Four mapping functions are explored in this section, i.e., linear regression, Kalman filter, Multi Layer Perceptrons (MLP) and Wavelet Neural Networks (WNN).

1) Linear Regression: Linear regression [12] finds a linear relationship between the inputs and output of the model. Based on (4), the inputs of the linear regression model are $p_j(t),\ p_j(t-1),\ldots,p_j(t-lag),\ \forall j=1,\ldots,J$, and its output is $\hat{P}_{tot}(t)$. Linear regression is very simple and easy to implement and hence, might be of great interest in industrial applications.

2) Kalman Filter: Kalman filter is a very useful tool that expresses the dynamics of a system by state-space equations. It takes into account uncertainties such as measurement and process noise [13]. Hence, it could be beneficial in invisible solar estimation and thus is implemented here. In the current study, an Auto Regressive (AR) model is used to model the linear relationship of $P_{tot}(t)$ to its previous lags. Moreover, measurement errors or any other source of uncertainty, such as passing clouds over the PV modules, are also considered. In general, each site has different characteristics and equipments and thus, the errors associated with their measurements would be different. However, in the current work there is no information available about the site's reliability. Hence, the uncertainty is assumed to be relative to the capacity of the sites, i.e., the sources of uncertainty are assumed to be the same across

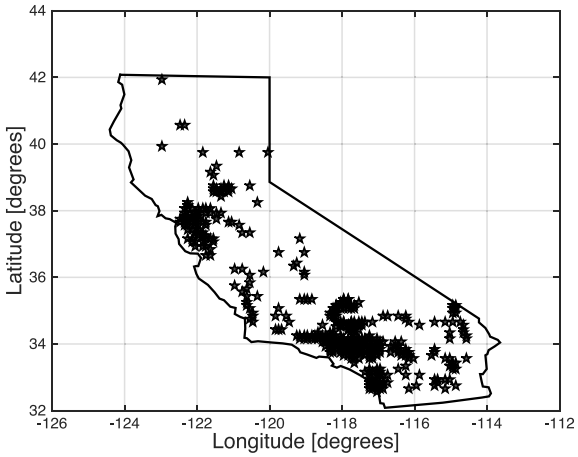

Fig. 3. The locations of 405 solar sites used for numerical simulations.

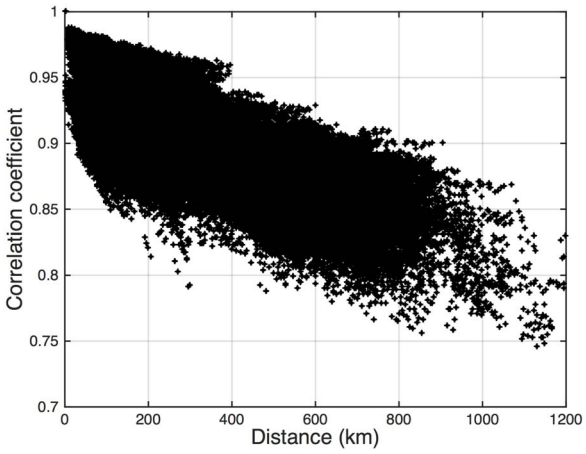

Fig. 4. Scatter plot of power generation correlation versus site physical distances for all sites.

the sites. It is also assumed that the uncertainty of the sites are not correlated to each other.

3) Multi Layer Perceptrons (MLP): Artificial neural networks are powerful tools for mapping a set of input variables to an output variable. MLP is one of the most frequently used engines in different applications and can be used to approximate any continuous nonlinear function to any accuracy level [14]. MLP has been implemented to approximate non-linear functions [30], and forecasting applications [15], [31]. The inputs of the MLP are data of selected J sites and the output of the model is the aggregated generation value of the total number of sites. Hence, the MLP has J inputs and only one output. The MLP is trained using the Levenberg-Marquardt (LM) learning algorithm [32]. The LM is known as an efficient and fast training algorithm for highly non-linear neural networks.

4) Wavelet Neural Networks (WNN): The WNN [15] is a type of neural network whose activation functions in the hidden layer are Wavelet functions. WNNs have been used in different forecasting applications [15], [23], [33], [34] and have shown promising results. Thus, a WNN with multi-dimensional Morlet Wavelet as the activation function is explored in this paper for the estimation stage.

III. Numerical Results

Ideally, proper numerical simulations for the proposed method should be done based on a very large set of roof-top solar module data. Collecting and processing such large set of data is still an ongoing effort by the authors. Observe that it is not a mandatory requirement for roof-top solar panel owners to collect and archive their power production data. Thus, finding relevant data in the public domain for expanding our numerical experiments has been challenging. We, however, believe that the presented numerical results provide good insight into the applicability of the proposed method.

For the purpose of proving the concept in this paper, we have used the NREL solar data [35], that is readily available. This set of data contains the estimated production of 405 PV-sites across California. Figure 3 shows the locations of the 405 sites. This number of sites is obviously far smaller than that

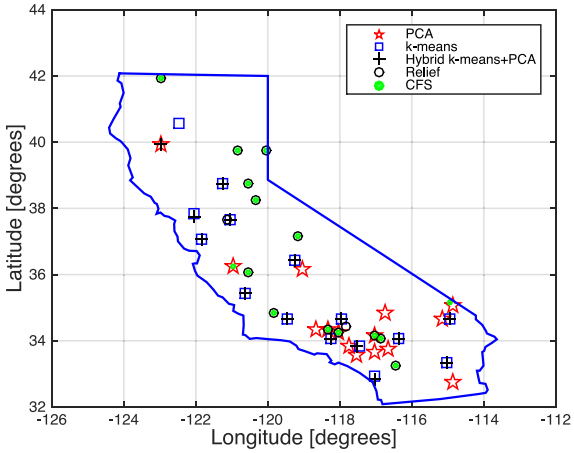

Fig. 5. Locations of the top 15 selected sites, i.e., $I = 405$ and $J = 15$.

of roof-top sites, yet we think the data could be used to prove the concept. We have downscaled the capacity of the sites to 1-50 kW to represent residential and commercial solar PV systems. We have used hourly production data. The daily total peak generation for the test data varies between 1,147.0 kW and 4,217.0 kW. This indicates that the data includes a variety of solar production regimes from high power during sunny days to lower power during cloudy days. Most of the sites are concentrated in western and southern California, where the major cities are located. The dataset includes production values for each individual site for a full year. We use four months worth of data for training the process, and use the remaining eight months for out-of-sample testing. Fig. 4 shows the scatter plot of the correlation coefficients among individual sites versus their physical distances. As expected, there are high correlations between closer sites, and thus, reducing the dimension of the data should not lead to loss of significant amount of information.

A. Data Dimension Reduction

We applied the five data dimension reduction techniques discussed in Section II-B to the hourly generation data for the 405 sites from the dataset. Figure 5 shows the location of

TABLE I
THE CAPACITY OF THE TOP 15 SELECTED SITES FOR DIFFERENT DIMENSION DATA REDUCTION TECHNIQUES IN KW. PA: THE PROPOSED HYBRID *k*-MEANS+PCA APPROACH

PCA		*k*-means		PA		Relief		CFS	
ID	Cap.	ID	Cap.	ID	Cap.	ID	Cap.	ID	Cap.
316	50	16	1.25	13	3.25	403	2.75	403	2.75
292	50	65	7.75	66	7.75	383	12.5	401	18.75
273	50	107	30.25	107	30.25	291	50	383	12.5
267	46.75	155	1	155	1	394	12.5	397	18.75
242	37.5	169	1.5	169	1.5	401	18.75	394	12.5
335	50	189	6.25	189	6.25	376	27.5	341	25
377	41	204	2.25	233	3.75	287	36.75	376	27.5
283	37.5	233	3.75	260	17.75	341	25	291	50
277	47.5	238	1.5	279	37.5	280	31.25	283	37.5
380	31.25	260	17.75	315	36.5	283	37.5	400	1.75
291	50	279	37.5	327	13	397	18.75	280	31.25
347	48.75	315	36.5	331	18.75	400	1.75	287	36.75
287	36.75	326	12.25	381	37.25	257	25	380	31.25
402	25	331	18.75	391	12.5	300	13	358	37.5
270	50	381	37.25	402	25	390	12.5	257	25

the selected sites by each of the five techniques when selecting the 15 most informative sites. While different methods select different sites, in general, the selected sites represent all concentrations of the sites across the sate. Among all the techniques, *k*-means shows better diversity in the placement of the selected sites. There is a reasonable similarity between the sites selected by Relief and CFS, which roots in the way these models select them. However, PCA selects the sites mainly in south where more generation capacity is available.

The hybrid *k*-means+PCA technique selects the sites that have the properties of the sites selected by both *k*-means Clustering and PCA techniques alone. As the figure shows, these sites are very close to those of the *k*-means approach. Since the combined approach still selects only one site for each cluster, between the two close sites the one with a higher PCA rank will be selected. These results are slightly different compared to those of the *k*-means technique. The value of threshold α has a key role in the performance of the hybrid *k*-means+PCA technique. Very small values of α lead to results closer to *k*-means Clustering, whereas large values of α move the results toward those of the PCA technique. The best α is found by trial and error, which is 0.01 that is used for the reported results in this paper.

Table I compares the capacity of the top 15 selected sites. As could be seen, PCA selects the sites with the maximum or close to maximum capacity. Other approaches select the sites from a variety of capacities. This means that the PCA is biased to the capacity of the sites. We also used the normalized data of the sites in the PCA approach, which in turn results in different ranking for the sites. However, the final estimation accuracy were not satisfactory and hence they are not reported here.

Table I also reveals that, for the selected top 15 sites, there is only one common site between PCA and the hybrid *k*-means+PCA approach. However, 10 sites are common between *k*-means Clustering and the hybrid *k*-means+PCA approach. Moreover, the capacity of the different sites are usually higher than those selected from *k*-means but none of them is 50 kW.

B. Estimating Invisible Solar PV Generation

We use the four mapping techniques, i.e., linear regression, Kalman filter, MLP, and WNN to estimate the hourly aggregated generation of the $I = 405$ sites in the dataset. For the MLP and WNN, respectively, 50 and 4 hidden neurones are found, by trial and error, to be the best configuration for this dataset. Furthermore, it has been found by trial and error that 24 hours was the best lag for the AR in the Kalman filter. In addition, in linear regression and the WNN, previous data associated with lags of $\{1, 2, 24, 25, 48\}$ hours are also included in the inputs. "Trial and error" here means that the authors have done the simulations using different configurations and the best results based on the eight months test data are reported here. On the other hand, these lags are not included in the Kalman filter and the MLP. Since the relation of aggregated generation with its past values is modeled by AR in the Kalman filter we did not include individual site's lagged values. Moreover, since including the lags will increase the number of the inputs, and thus complexity, of the MLP, the final results found to be worse compared to only using data of time *t*.

We use average daily Root Mean Squared Error (DRMSE) values to measure the accuracy of the proposed approach, which is defined as follows:

$$DRMSE = \frac{1}{ND}\sum_{d=1}^{ND}\left\{\sqrt{\frac{1}{N_d}\sum_{t\in T_d}\left[P_{tot}(t) - \hat{P}_{tot}(t)\right]^2}\right\} \quad (8)$$

where, *ND* is total number of test days, N_d is the number of daylight hours of day d, and T_d is the set of all daylight time stamps on day d. The RMSE values are calculated for each day only for daylight hours and then their averages over the 8-month test period are reported in Table II. The smallest DRMSE among each estimation model is highlighted in gray. Moreover, in the table the results for the best combination of data reduction approach and estimation function for each number of selected sites is shown in bold font.

From Table II, observe that the *k*-means Clustering algorithm generally serves better than other individual data reduction approaches for all four estimation engines. This means that locational data is very valuable in order to select the best candidate sites. However, in some cases, such as the case with 10 sites in the linear model, PCA has less estimation error compared to the *k*-means sites. In addition, Relief and CFS are the worst approaches to select the candidate sites and the final estimation errors are high for all of the four models.

The other observation from Table II is that the proposed hybrid *k*-means+PCA data reduction method shows significant improvements in the final estimation results in most of the cases compared to the *k*-means Clustering and PCA approaches. In fact, in most of the cases the proposed hybrid *k*-means+PCA approach outperforms all four individual data reduction techniques. Although we expect less DRMSE by selecting more sites, this does not always hold, specially when using *k*-means approach. The reason is that in some cases, the centroid site does not have a good quality output behavior to be able to represent the neighboring sites. This leads to less accurate results. The proposed hybrid *k*-means+PCA

TABLE II
THE DRMSE IN, KW, OF DIFFERENT MODELS FOR DIFFERENT NUMBER OF SELECTED SITES. KM: *k*-MEANS; PA: THE PROPOSED HYBRID *k*-MEANS+PCA APPROACH; REL: RELIEF

No. of Sites	Linear Regression					Kalman Filter					MLP					WNN				
	PCA	KM	PA	Rel	CFS	PCA	KM	PA	Rel	CFS	PCA	KM	PA	Rel	CFS	PCA	KM	PA	Rel	CFS
2	422	368	553	789	878	396	342	585	862	1088	404	**227**	475	810	785	372	311	565	810	935
3	237	268	264	532	788	301	278	293	514	903	265	**193**	217	574	640	293	278	280	570	785
4	227	242	227	531	781	287	317	279	524	960	241	**182**	226	468	654	258	221	231	469	783
5	202	**147**	150	532	770	285	182	163	559	949	237	206	156	450	609	224	156	157	499	718
6	201	148	**144**	531	623	289	179	162	562	777	218	177	152	468	895	194	171	148	540	834
7	146	109	**106**	496	616	228	171	123	511	732	187	132	137	472	479	155	109	108	513	664
8	147	112	**101**	495	521	247	166	140	512	583	157	134	135	404	439	143	121	105	461	567
9	142	209	**95**	480	515	249	196	135	492	545	154	154	111	403	379	166	190	99	458	456
10	132	210	**95**	489	345	245	195	130	483	543	142	153	110	383	333	136	162	97	412	314
15	100	99	**85**	265	254	211	136	129	472	449	113	112	109	306	275	110	93	85	315	261
20	92	80	**72**	231	191	210	132	132	404	416	120	98	96	244	268	100	81	73	224	186
25	84	75	**71**	167	169	194	117	99	363	399	107	97	94	195	204	92	76	72	169	185
30	84	73	**65**	150	159	173	111	107	350	363	159	92	78	180	199	86	74	67	153	179

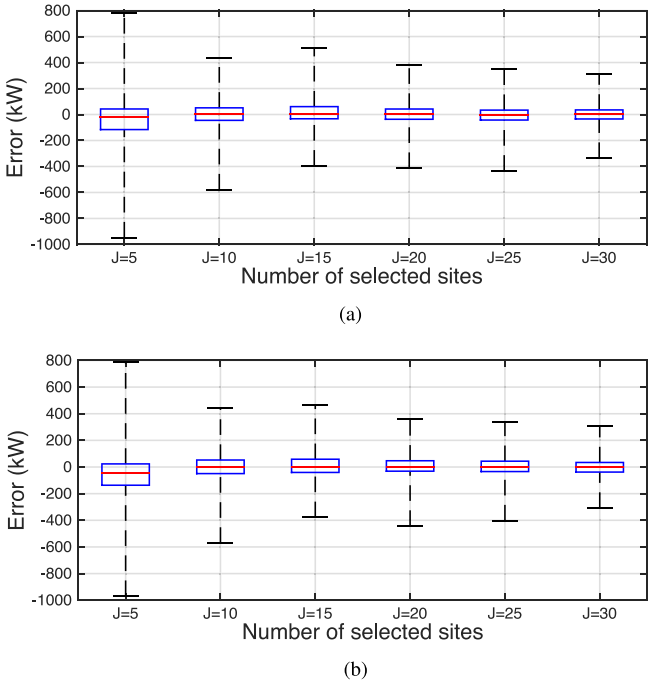

Fig. 6. Box plot of estimation errors using the proposed hybrid *k*-means+PCA approach: a) Linear regression estimation function; b) WNN estimation function.

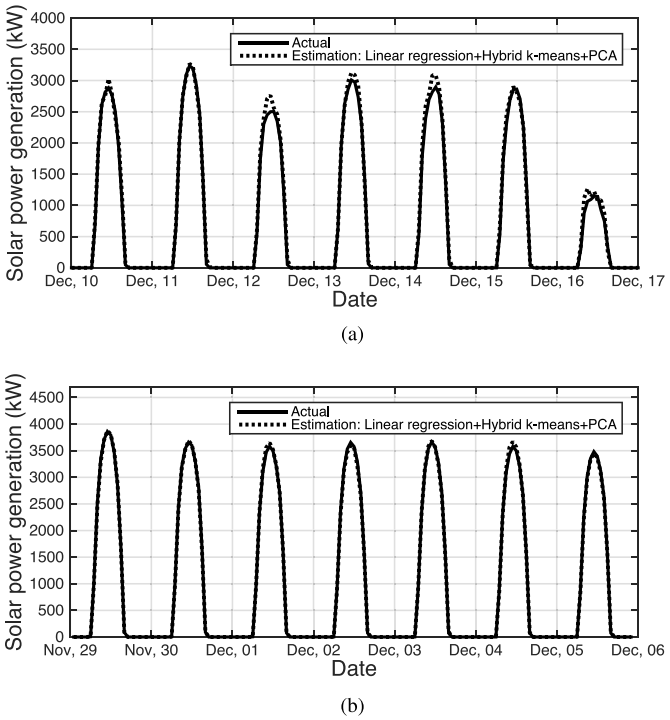

Fig. 7. One week of estimations and actual values for $J = 15$ using linear regression equipped with the proposed hybrid *k*-means+PCA approach: a) Weak estimations; b) Good estimations.

approach solves the issue of bad site selections for *k*-means, e.g., selection of 9-10 sites. Moreover, the proposed hybrid *k*-means+PCA data dimension reduction approach leads to lower errors compared to the other approaches. From the table, and for this particular dataset, the linear regression model along with the proposed hybrid *k*-means+PCA data dimension reduction method has outperformed other methods from a DRMSE viewpoint.

Figure 6 depicts box plot of the estimation errors during daylight hours for the two best models, i.e., hybrid *k*-means+PCA approach with linear regression in the top and WNN on the bottom. As both Fig. 6 and Table II reveal, the more sites used, the better the results could be. However, the rate of estimation accuracy improvement decreases as the number of the sites increases. Figure 6 clearly shows that the difference between the performance of the two best models is marginal for $J = 15$ and above. This means that although using more sites leads to better results, there needs to be a tradeoff between the number of selected sites and the desired accuracy of the estimations.

Figure 7 depicts the estimation results, along with the actual values, for the linear regression function and the proposed hybrid *k*means+PCA data reduction approach for $J = 15$. In the top plot, the method estimates the total power generation reasonably well for the high-production days. However, for the cloudy days that seem to have had less sunshine, the model misses the actual values to some extent. This is mainly because no weather-related information is included in the simulations

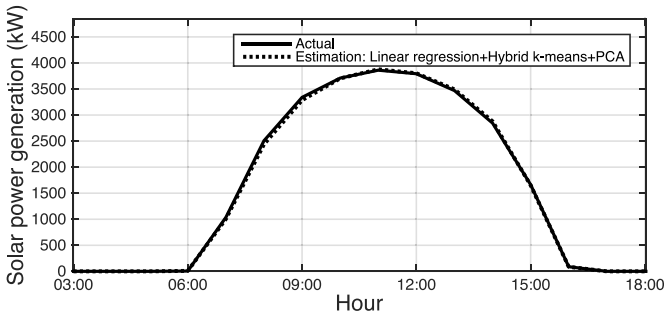

Fig. 8. Estimations and actual values of one day for $J = 15$ using linear regression equipped with the proposed hybrid k-means+PCA approach: a) Sunny day of November 29; b) Cloudy day of December 16.

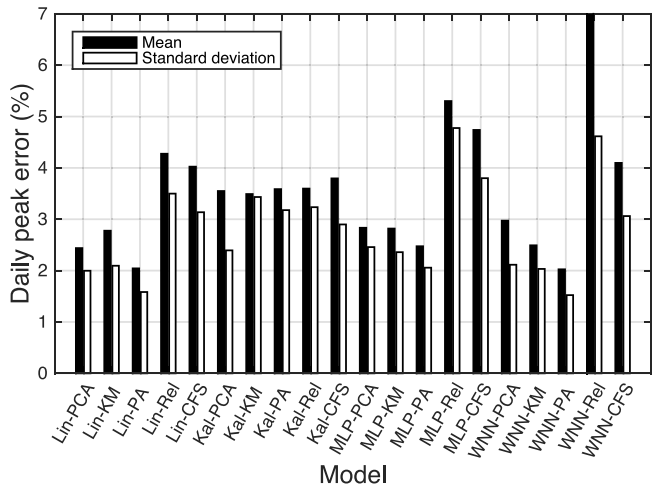

Fig. 9. Daily peak estimation error for $J = 15$ candidate sites. KM: k-means; PA: the proposed hybrid k-means+PCA approach; Rel: Relief; Lin: linear regression; Kal: Kalman filter.

and the models are purely built based on past power values. As the bottom plot shows, better estimations are made for the days that have a consistent power production pattern. Figure 8 shows the estimation results and actual values for a sunny day at the top and a cloudy day on the bottom. As the figure shows, the model performs very well on sunny days but has higher errors of up to 200 kW on the cloudy day.

Daily peak is an important measure in power system operation. Daily solar power generation peak error results in wrong calculation of the daily net load valley, which in turn could jeopardize the security of the system and impose extra costs.

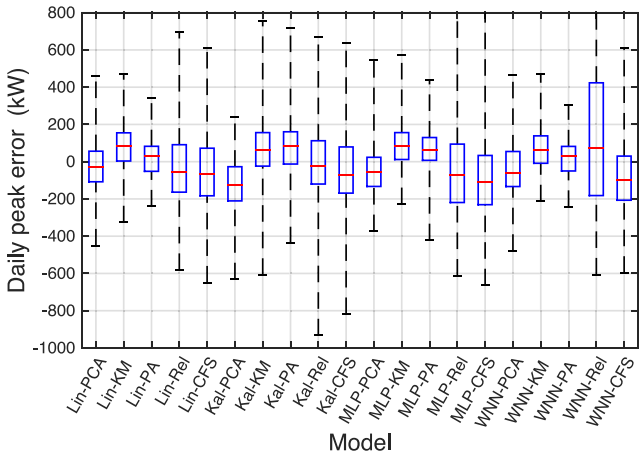

Fig. 10. Box plot of daily peak estimation error for $J = 15$ candidate sites. KM: k-means; PA: the proposed hybrid k-means+PCA approach; Rel: Relief; Lin: linear regression; Kal: Kalman filter.

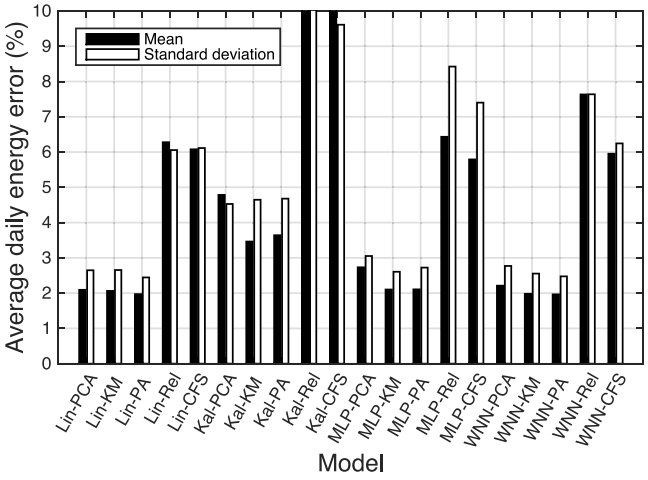

Fig. 11. Daily energy estimation error for $J = 15$ candidate sites. KM: k-means; PA: the proposed hybrid k-means+PCA approach; Rel: Relief; Lin: linear regression; Kal: Kalman filter.

Fig. 9 summarizes the average daily peak estimation errors for different approaches over the estimation period, i.e., 8 months, for $J = 15$. From this figure, the proposed hybrid k-means+PCA approach performs better than other data reduction techniques for both the average and standard deviation of the errors. Another observation is that although MLP and WNN perform relatively close to the other two strategies using PCA and k-means, their results for the Relief and CFS are worse than others. This means that neural networks could be more sensitive to data reduction methods for daily peak estimation. In addition, Fig. 10 presents box plot of the daily peak estimation error for different approaches with $J = 15$. This figure also confirms that the hybrid k-means+PCA approach performs better than other data reduction techniques and the best results are associated with linear regression and WNN estimation functions.

We also analyzed the daily energy estimation errors for all the combinations. Fig. 11 shows the related daily energy estimation errors for different methods for

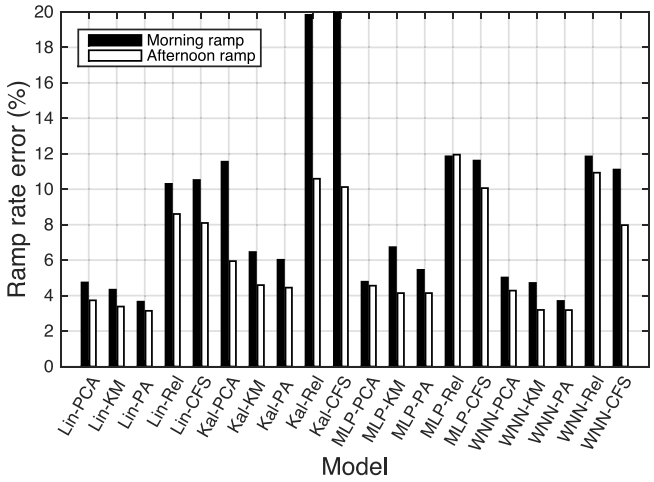

Fig. 12. Average daily ramp rate error for $J = 15$ candidate sites. KM: k-means; PA: the proposed hybrid k-means+PCA approach; Rel: Relief; Lin: linear regression; Kal: Kalman filter.

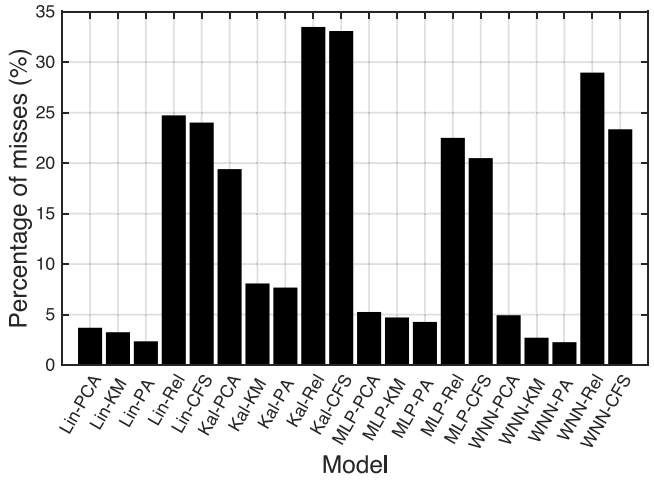

Fig. 13. Number of hours that estimation is out of the 5% actual interval for $J = 15$ candidate sites. KM: k-means; PA: the proposed hybrid k-means+PCA approach; Rel: Relief; Lin: linear regression; Kal: Kalman filter.

15 selected sites. Similar to the previous case, the proposed hybrid k-means+PCA data reduction is the best of the models. However, here the improvements compared to the k-means or PCA is marginal. Moreover, the difference between Relief and CFS compared to others is very high and they performed poorly in this regard. The results also show that the proposed hybrid k-means+PCA data reduction technique along with linear regression or WNN outperforms other methods.

Ramp rate, which is the changes of power generation during a specified time interval is a critical parameter in system operation. In this work, we have defined morning ramp rate as the average hourly change in the aggregated generation from 7 to 10 am. Similarly, afternoon ramp rate is defined as the average hourly change during 3 pm to 6 pm. The mean and standard deviation of average morning and afternoon ramp rate errors of different models is presented in Fig. 12. These results also show that the proposed hybrid k-means+PCA data

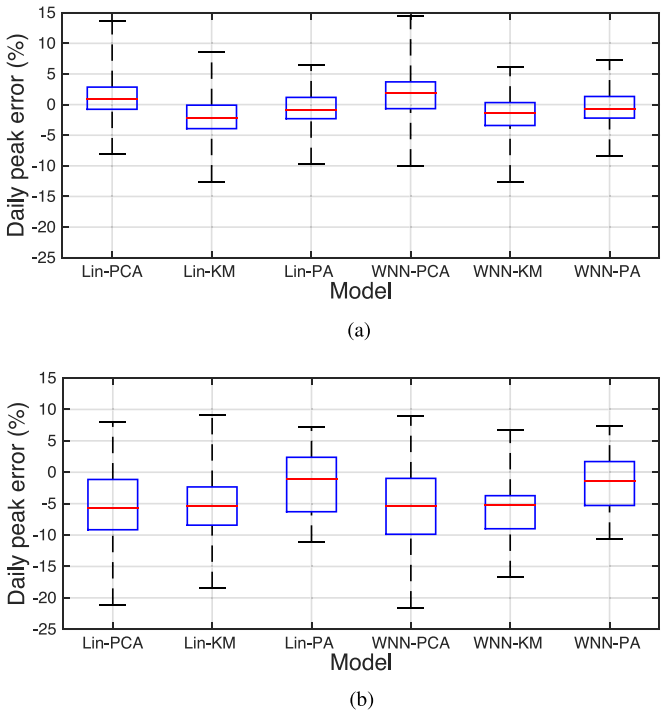

Fig. 14. Box plot of daily peak estimation error for $J = 15$ candidate sites: a) High generation days; b) Low generation days. KM: k-means; PA: the proposed hybrid k-means+PCA approach; Lin: linear regression.

reduction methodology along with linear regression or WNN outperforms other methods.

We have also analyzed the results to see how often the estimated total power generation values differ from the actual values by less than 5% of the peak power generation, and the results are presented in Fig. 13 for $J = 15$. Observe that the proposed hybrid k-means+PCA approach outperforms all of the other data reduction techniques by missing the margin in about 2.2% of the time in both linear regression and the WNN estimation functions.

We have also analyzed the results separately for high and low generation days to see how the proposed approach works on sunny and cloudy days. In order to do this, we have considered days that have daily aggregated peak generation of less than 3,000 kW, which is 70% of the annual peak generation, to be low generation days. Out of 245 test days, 27 days were considered to have low generation. The 70% threshold is chosen to have enough 'low' generation test days for reasonable conclusions. Note that because of the generally good sunshine regime in California most of the test days in our case study have a high daily peak solar power generation. Only 2 days and 9 days have the daily peak of less than 2000 kW and 2500 kW, respectively. Figures 14 and 15 show box plot of daily peak and energy estimation errors of the two day types tor $J = 15$ for the best combination of models, respectively. Observe from the figures that in all of the cases the proposed hybrid k-means+PCA data dimension reduction approach outperforms the other approaches for both high and low generation days and its results are more robust compared to the k-means and PCA approaches. Both figures also show that the errors are generally higher for cloudy days

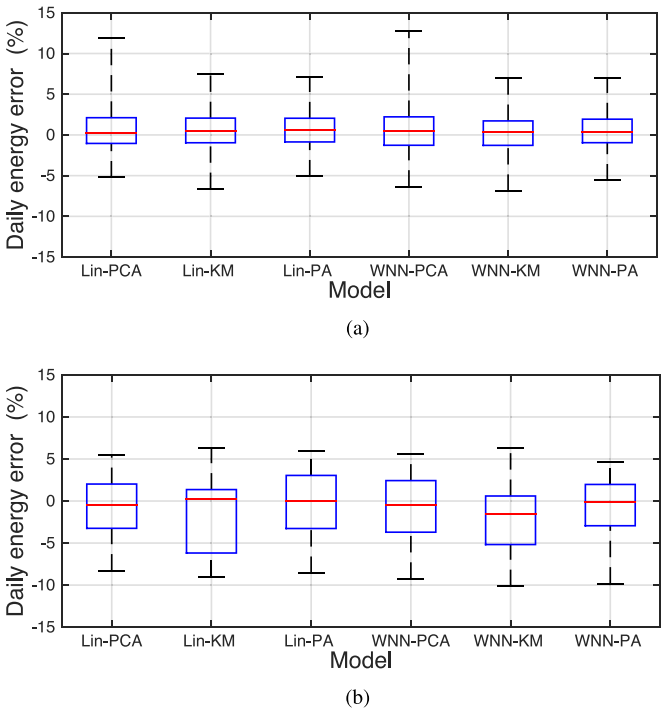

Fig. 15. Box plot of daily energy estimation error for $J = 15$ candidate sites: a) High generation days; b) Low generation days. KM: *k*-means; PA: the proposed hybrid *k*-means+PCA approach; Lin: linear regression.

compared to sunny days. However, results from the proposed hybrid *k*-means+PCA data dimension reduction approach are more stable compared to the other approaches.

IV. Conclusion

This paper proposes a methodology to estimate total generation from invisible solar PV sites. The proposed method has two main components, i.e., a data dimension reduction engine and a mapping function. The data dimension reduction engine selects a small subset of most informative sites from the large set of all available sites. The mapping function estimates the total power generation of all sites based on the data from the selected subset of sites. We use PCA, *k*-means Clustering, Relief, and CFS for data dimension reduction. We also propose a new hybrid *k*-means+PCA data reduction approach that combines *k*-means Clustering with PCA. The proposed algorithm is evaluated using the solar PV generation data of California's power system with a total of 405 sites. We provide discussions on average daily root mean squared estimation error, daily peak estimation error, daily energy estimation error, and ramp rate estimation error.

The results showed that the proposed framework is capable of estimating invisible solar generation with a good accuracy. In addition, the proposed hybrid *k*-means+PCA approach is found to outperform all of other data dimension reduction techniques. Although WNN showed similar or slightly better results compared to linear regression, considering the simplicity of the latter, it may better suite practical applications. In addition, the combination of linear regression and the proposed hybrid *k*-means+PCA dimension reduction method was found to generally outperform other combinations with respect to estimating the daily peak, total generated energy, and morning and afternoon ramps.

References

[1] *Go Solar California*. [Online]. Available: http://gosolarcalifornia.org, accessed Aug. 26, 2015.

[2] *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.

[3] *IESO, Independent Electricity System Operator: FIT Program*. [Online]. Available: http://fit.powerauthority.on.ca/fit-program, accessed Apr. 9, 2015.

[4] *Tesla Motors*. [Online]. Available: http://www.teslamotors.com, accessed Nov. 1, 2015.

[5] *Powerwall: Tesla Home Battery*. [Online]. Available: http://www.teslamotors.com/powerwall, accessed May 8, 2015.

[6] J. S. John. *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, accessed Oct. 2013.

[7] *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.

[8] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[9] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques* (Data Management Systems), 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[10] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. 9th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 1992, pp. 249–256.

[11] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1999.

[12] X. Yan and X. G. Su, *Linear Regression Analysis: Theory and Computing*. Singapore: World Scientific, 2009.

[13] J. D. Hamilton, *Time Series Analysis*, 2nd ed. Princeton, NJ, USA: Princeton Univ. Press, 1994.

[14] E. K. Blum and L. K. Li, "Approximation theory and feedforward networks," *Neural Netw.*, vol. 4, no. 4, pp. 511–515, 1991.

[15] N. M. Pindoriya, S. N. Singh, and S. K. Singh, "An adaptive wavelet neural network-based energy price forecasting in electricity markets," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1423–1432, Aug. 2008.

[16] *Enphase*. [Online]. Available: https://enphase.com, accessed Aug. 13, 2015.

[17] L. Xie, Y. Chen, and P. R. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2784–2794, Nov. 2014.

[18] J. Zhang, Y. Zhang, Z. Wang, and J. Ma, "PCA fault feature extraction in complex electric power systems," *Adv. Elect. Comput. Eng.*, vol. 10, no. 3, pp. 102–107, 2010.

[19] F. Vallee, G. Brunieau, M. Pirlot, O. Deblecker, and J. Lobry, "Optimal wind clustering methodology for adequacy evaluation in system generation studies using nonsequential Monte Carlo simulation," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2173–2184, Nov. 2011.

[20] N. Amjady, A. Daraeepour, and F. Keynia, "Day-ahead electricity price forecasting by modified relief algorithm and hybrid neural network," *IET Gener. Transm. Distrib.*, vol. 4, no. 3, pp. 432–444, Mar. 2010.

[21] I. Koprinska, M. Rana, and V. G. Agelidis, "Correlation and instance based feature selection for electricity load forecasting," *Knowl.-Based Syst.*, vol. 82, pp. 29–40, Jul. 2015.

[22] S. A. Bouhamed, I. K. Kallel, D. S. Masmoudi, and B. Solaiman, "Feature selection in possibilistic modeling," *Pattern Recognit.*, vol. 48, pp. 3627–3640, Nov. 2015.

[23] Y. Chen *et al.*, "Short-term load forecasting: Similar day-based wavelet neural networks," *IEEE Trans. Power Syst.*, vol. 25, no. 1, pp. 322–330, Feb. 2010.

[24] D. Huang, H. Zareipour, W. D. Rosehart, and N. Amjady, "Data mining for electricity price classification and the application to demand-side management," *IEEE Trans. Smart Grid*, vol. 3, no. 2, pp. 808–817, Jun. 2012.

[25] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," *J. Mach. Learn. Res.*, Jan. 2008. [Online]. Available: http://www.iai.uni-bonn.de/∼jz/dimensionality_reduction_a_comparative_review.pdf, accessed Nov. 1, 2015.
[26] I. K. Fodor, "A survey of dimension reduction techniques," Center Appl. Sci. Comput., Lawrence Livermore Nat. Lab., Livermore, CA, USA, Tech. Rep. UCRL-ID-148494, 2002, pp. 1–18.
[27] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," *Mach. Learn.*, vol. 75, no. 2, pp. 245–248, Jan. 2009.
[28] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Proc. Eur. Conf. Mach. Learn.*, Catania, Italy, 1994, pp. 171–182.
[29] *Weka 3: Data Mining Software in Java*. [Online]. Available: http://www.cs.waikato.ac.nz/∼ml/weka/, accessed Nov. 1, 2015.
[30] A. M. Patrikar, "Approximating Gaussian mixture model or radial basis function network with multilayer perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1161–1166, Jul. 2013.
[31] R. Ak, O. Fink, and E. Zio, "Two machine learning approaches for short-term wind speed time-series prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.
[32] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 989–993, Nov. 1994.
[33] H. Shaker, H. Chitsaz, H. Zareipour, and D. Wood, "On comparison of two strategies in net demand forecasting using wavelet neural network," in *Proc. North Amer. Power Symp. (NAPS)*, Pullman, WA, USA, Sep. 2014, pp. 1–6.
[34] L. J. Ricalde, G. A. Catzin, A. Y. Alanis, and E. N. Sanchez, "Higher order wavelet neural networks with Kalman learning for wind speed forecasting," in *Proc. IEEE Symp. Comput. Intell. Appl. Smart Grid (CIASG)*, Paris, France, Apr. 2011, pp. 1–6.
[35] *NREL Solar Power Data for Integration Studies*. [Online]. Available: http://www.nrel.gov/electricity/transmission/solar_integration_methodology.html, accessed Jan. 14, 2014.

Hamid Shaker (S'12) received the B.Sc. degree from the Isfahan University of Technology, Isfahan, Iran, in 2007, and the M.Sc. degree from the Sharif University of Technology, Tehran, Iran, in 2009, both in electrical engineering. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Calgary, Calgary, AB, Canada. His current 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 (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, U.K., in 1980. He has been a Professor and NSERC/ENMAX Industrial Research Chair in renewable energy with the Department of Mechanical and Manufacturing Engineering, University of Calgary, Calgary, AB, Canada, since 2010. His current research interests include small wind turbines and other forms of renewable energy, including resource assessment and forecasting.