Analyzing big time series data in solar engineering using features and PCA

Dazhi Yang[a,*], Zibo Dong[b], Li Hong I. Lim[c], Licheng Liu[d]

[a]*Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore*
[b]*Solar Energy Research Institute of Singapore, National University of Singapore, Singapore*
[c]*Department of Electronic Systems, University of Glasgow, UK*
[d]*Saferay Pte. Ltd., Singapore*

Abstract

In solar engineering, we encounter big time series data such as the satellite-derived irradiance data and string-level measurements from a utility-scale photovoltaic (PV) system. While storing and hosting big data are certainly possible using today's data storage technology, it is challenging to effectively and efficiently visualize and analyze the data. We consider a data analytics algorithm to mitigate some of these challenges in this work. The algorithm computes a set of generic and/or application-specific features to characterize the time series, and subsequently uses principal component analysis to project these features onto a two-dimensional space. As each time series can be represented by features, it can be treated as a single data point in the feature space, allowing many operations to become more amenable. Three applications are discussed within the overall framework, namely (1) the PV system type identification, (2) monitoring network design, and (3) anomalous string detection. The proposed framework can be easily translated to many other solar engineer applications.

Keywords: Principal component analysis, Time series features, Solar irradiance, Characterization

1. Introduction

Many solar engineering datasets, such as high-resolution satellite-derived irradiance data (e.g., Nikitidou et al., 2015), power output data from hundreds of photovoltaic (PV) plants in an area (e.g., Yang et al., 2017) and module-level measurements from a PV plant (e.g., Guerriero et al., 2016), align well with the HACE theorem[1] proposed by Wu et al. (2014) that characterizes big data. One of the main challenges of processing these raw datasets is the high noise and irrelevant information embedded. Moreover, visualization and analysis through operating directly on the

*Corresponding author. Tel.: +65 9159 0888.
Email address: yangdazhi.nus@gmail.com; yangdz@simtech.a-star.edu.sg (Dazhi Yang)

[1]Big Data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data (Wu et al., 2014).

raw datasets can be ineffective. On this point, the Pareto principle, better-known as the 80/20 rule, commonly applies: researchers and solar engineers often spend most of their time collecting, cleaning, filtering, reducing and formatting the data. In this paper, a data analytics algorithm is used to overcome some of the aforementioned challenges. We will look into a class of applications which involve big time series data, or more specifically, solar irradiance and other related forms.

A time series is a collection of observations taken sequentially in time; this definition provides a natural grouping for the data. Instead of viewing the data points as individual entities, we can view time series as *entities*. Once this seemingly trivial statement is understood, much convenience can be added to data handling and analytics. Traditionally, to reduce the complexity in time series data, we often shorten each time series, but preserve the number of entities. For example, satellite-derived irradiance data can be considered as time series of lattice processes. As the data usually span decades, some reduced form, such as a typical meteorological year (TMY) file, can be useful. Composition of the TMY data typify conditions at a particular site over a longer period of time, i.e., 10 to 30 years. For computer simulations of solar energy conversion systems and building systems to facilitate performance comparisons of different designs, this type of reduced dataset is sufficient (Wilcox and Marion, 2008). Our approach of representing raw time series is similar to the construction of TMY datasets.

The core concept is rather simple: each time series is treated as an individual entity which can be characterized by a set of generic or application-specific features. This step dramatically reduces the dimension of the data, i.e., from hundreds of samples in a time series to a few descriptive features. As each time series can now be treated as a single data point in the feature space, many operations become more amenable in that feature space. Furthermore, it is also easier to visualize big time series data in the feature space as compared to the traditional time series visualization methods such as the spaghetti plot and horizon plot, which are informative but not very scalable. We illustrate these points with a toy example.

1.1. A toy example

Let us consider the problem of detecting faulty strings using commonly available data from a PV plant. Suppose we represent each string-level output current time series with a single feature, namely the mean value over a period of time, and plot it on the real line, the faulty string could be detected by locating the outliers in that one-dimensional feature space, as illustrated in Figure 1. While the single feature approach may allow us to detect the faulty strings, it is difficult to isolate the fault type. The decrease in output current is a shared observation for several different fault types (see Table 2 in Chine et al., 2016). If a second feature is added, namely, the mean output voltage over that period of time, the faulty strings can now be represented in a two-dimensional space. Since the voltage of the faulty strings can increase, decrease or remain constant, corresponding to different faults, combining two features would better identify fault types.

The above idea can immediately be expanded to a feature space with p dimensions, with the additional features being, e.g., mean short circuit current, mean open circuit voltage and number of maximum power point; these features were used in Chine et al. (2016). Within the narrow premise of this toy example, more features imply better isolation of faults. In the ideal case, the described approach could circumvent the tedious string-by-string fault check, which often involves complex procedure and flow chart.

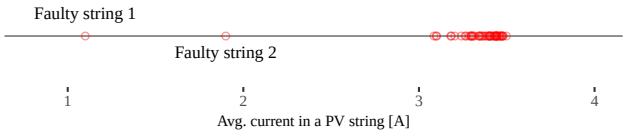

Figure 1: An illustrative example of PV string fault detection in a one-dimensional feature space.

The one-dimensional case displayed in Figure 1 provides excellent visualization of multiple time series. It is not difficult to imagine such plots in a two- or three-dimensional space. When $p > 3$, the scatter plot can still be visualized by performing principal component analysis (PCA) on the features. PCA uses an orthogonal transformation to convert possibly correlated features to linearly uncorrelated variables, known as principal components (PCs). As the first few PCs often contain most variation, it is common to plot the data points – recall each data point represents a time series in our case – in a new low-dimensional space constructed by the directions of the first few PCs. When PCA is considered, its companion algorithms, such as the k-means clustering, α-hull and high density region, can be then applied to solve a variety of problems, and thus make the data analytics algorithm very versatile.

1.2. Applications

We study three applications in this work, namely, (1) PV system type identification, (2) monitoring network design and (3) anomalous PV string detection. We note that all three applications are well-studied in the literature (the literature review will distribute to respective sections), however, the merit of the present work goes to the new point of view on data handling. In clustering problem like the first two applications, the k-means algorithm will be used together with PCA. Unlike other alternatives, this approach does not cluster raw point values using a distance metric, rather it clusters based on global features extracted from each time series. The third application is in line with the toy example above. A two-dimensional outlier detection algorithm, the α-hull algorithm, will be applied to the result of PCA. This is also distinct from most outlier detection studies in the literature, where outliers are identified within one time series or based on statistical rules. Besides the these applications, there are many other applications that could potentially benefit from the analytics algorithm. We briefly enumerate several other applications in Appendix E.

2. Principal component analysis and biplot

For a *centered* dataset $\boldsymbol{X}$, an $n \times p$ matrix, where n is the number of time series (observation, each time series is considered as one observation) and p is the number of time series features (variable), PCA computes the most meaningful[2] *basis* to re-express $\boldsymbol{X}$. If $\boldsymbol{Z}$ is the re-represented data, the above statement can be written as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}$, where $\boldsymbol{A}$ is an $p \times p$ matrix and its columns are a set of basis vectors for representing of columns of $\boldsymbol{X}$.

[2]For a detailed discussion on the motivation for PCA, and what should be considered as "most meaningful", we refer the readers to Shlens (2003).

PCA assumes all basis vectors are orthonormal. It first selects a normalized direction in p-dimensional space along which the variance in $\boldsymbol{X}$ is maximized; this basis vector is denoted as $\boldsymbol{a}_1$. In other words, we maximize $\mathbb{V}(\boldsymbol{a}_1^\top \boldsymbol{x})$, where $\boldsymbol{x}$ is vector of p random variables (p time series features in this case). Since the maximum will not be achieved with finite $\boldsymbol{a}_1$, a normalization constraint is imposed, namely, $\boldsymbol{a}_1^\top \boldsymbol{a}_1 = 1$. The subsequent direction is again selected based on the maximum variance criterion, however, due to the orthonormal assumption, the choice is limited to the directions that are perpendicular to $\boldsymbol{a}_1$. The procedure continues until p directions are selected. Thus $\boldsymbol{a}_k^\top \boldsymbol{x}$ is defined as the kth sample principal components and $z_{ik} = \boldsymbol{a}_k^\top \boldsymbol{x}_i$ is the score for the ith observation on the kth PC.

2.1. Solving PCA with eigendecomposition

As the goal of PCA is to reduce redundancy, it is desired that each variable co-varies as little as possible with other variables. In other words, we aim to diagnolize the covariance matrix of the re-represented data. Let $\boldsymbol{S_Z}$ be the covariance matrix of $\boldsymbol{Z}$, i.e.,

$$\boldsymbol{S_Z} = \frac{1}{n-1}\boldsymbol{Z}^\top \boldsymbol{Z}, \tag{1}$$

we have

$$\begin{aligned}
\boldsymbol{S_Z} &= \frac{1}{n-1}(\boldsymbol{XA})^\top(\boldsymbol{XA}) \\
&= \frac{1}{n-1}\boldsymbol{A}^\top(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{A} \\
&= \frac{1}{n-1}\boldsymbol{A}^\top(\boldsymbol{EDE}^{-1})\boldsymbol{A},
\end{aligned} \tag{2}$$

where $\boldsymbol{E}$ is a matrix of eigenvectors of $\boldsymbol{X}^\top \boldsymbol{X}$ arranged as columns and $\boldsymbol{D}$ is a diagonal matrix. If we let $\boldsymbol{A} \equiv \boldsymbol{E}$, the covariance matrix

$$\begin{aligned}
\boldsymbol{S_Z} &= \frac{1}{n-1}\boldsymbol{A}^\top(\boldsymbol{ADA}^{-1})\boldsymbol{A} \\
&= \frac{1}{n-1}(\boldsymbol{A}^{-1}\boldsymbol{A})\boldsymbol{D}(\boldsymbol{A}^{-1}\boldsymbol{A}) \\
&= \frac{1}{n-1}\boldsymbol{D}
\end{aligned} \tag{3}$$

can be diagnolized (note that $\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$ when $\boldsymbol{A}$ is orthogonal). This was the goal for PCA. The eigenvectors can be found via eigendecomposition. Alternatively, a more mathematically involved approach to solve PCA is through singular value decomposition (SVD). One advantage of using SVD is that it can handle the situation where there are more dimensions than samples. However, this is rarely the case when big time series data is considered.

The derivation shown above are based on the eigenvectors and eigenvalues of the covariance matrix. However, when the variables in the centered dataset vary by orders of magnitude, performing PCA with this data will lead to large loadings[3] for variables with high variance, which is

[3]Eigenvectors are cosines of rotation of variables into components. Loadings are eigenvectors normalized to respective eigenvalues, i.e., $loading = eigenvectors \cdot \sqrt{eigenvalues}$.

undesirable. It is therefore appropriate to perform scaling before PCA; this is achieved by dividing each variable by its standard deviation.

There are various statistical packages which provide implementations of PCA. For instance two popular functions in software R, namely, `princomp` and `prcomp`, which compute principal components using eigendecomposition and SVD, respectively. In this paper, we use `princomp` throughout and the supplementary material, i.e., the R code, is provided in Appendix A.

2.2. Biplot

A biplot represents both the observations and variables of a matrix of multivariate data on the same plot. It uses points to represent the scores of the observations on the principal components, and uses vectors to represent the coefficients of the variables on the principal components. Superimposing the observations and variables provides additional insights about relationships between them not available in either individual plot (Jolliffe, 2002).

Recall that PCA splits covariance or correlation matrix into a scale part (eigenvalues) and a direction part (eigenvectors), plotting loadings instead of eigenvectors makes them comparable by magnitude with the covariance or correlation observed between the variables. In fact, loadings are the covariances or correlations between the original variables and the unit-scaled components. This point will be reiterated in later sections of this paper.

Biplots are informative; we will thus interpret them in the following three sections. At this stage, several interpretations are summarized:

1. Biplots are scatter plots, the points in a biplot can therefore be interpreted the same way: closer points correspond to observations that have similar scores on the PCs. This interpretation is useful for clustering applications;
2. Projection of points onto a vector gives original values of that variable. By examining points along the particular direction of a vector (and its opposite direction as well), samples with anomalous values on what the variable measures can be identified;
3. Angle between two vectors denotes their correlation. Vectors that point in the same direction correspond to variables that have similar response profiles;
4. The apparent length of a vector gives an idea about the variance of that variable. This can be used to conclude the importance of a variable during clustering.

3. Applications A: PV system type identification (generic time series features)

The first, and arguably the simplest, application is PV system type identification. In particular, we are interested in identifying whether a PV system has a fixed orientation or single-axis tracking; these two types of systems are more utilized than duel-axis tracking systems due to their better cost, reliability and energy production trade-off (Mousazadeh et al., 2009; Nann, 1990).

The data used in this application comes from the western wind and solar integration study (WWSIS). WWSIS is a three-phase project (GE Energy, 2010; Lew et al., 2013; Miller et al., 2014) conducted by the National Renewable Energy Laboratory (NREL) to explore the operational impact of high renewable penetration into an electricity grid. The sub-hour solar irradiance data in WWSIS were synthetically generated using the algorithm developed by Hummon et al. (2012).

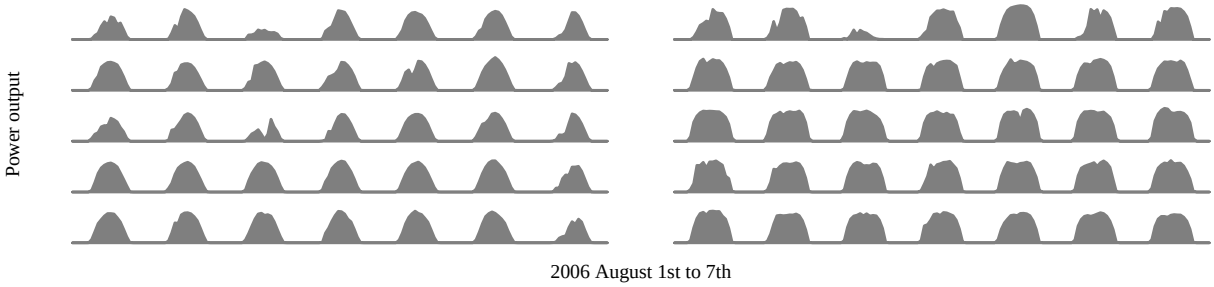

Figure 2: Sample time series from the WWSIS dataset during 2006 August 1 to 7. Power output from PV systems with trackers (right column) has a flatter top as compared to that from systems with fixed orientations (left column).

The irradiance was then converted to solar power output through the System Advisor Model. The full dataset contains approximately 6000 PV plants of different size in western US locations, however, for demonstration purposes, only data from 405 plants in California are used. The data is first normalized by dividing the series with its respective system size.

Figure 2 shows some samples of the normalized time series from the WWSIS dataset (five random samples from each type of systems). It can be seen that the power output from those PV systems with trackers (plotted in the column on the right) has a flatter top, due to the DNI gain by the sun-tracking panels. This effect is most apparent in the late morning and early afternoon hours. Although identifying PV systems types by visually inspecting the time series transient is easy, the method is not scalable when thousands or more series need to be identified. The proposed framework can be useful in this application. Due to the simplicity of this application, using only *generic time series features* would suffice.

Generic time series features refer to the statistics (such as mean, variance and autocorrelation) and results of simple counting (such as number of times a series crosses its mean and length of a time series) which can be applied to most, if not all, time series. Hyndman et al. (2015) consolidated a total of 15 generic time series features; these features are listed verbatim in Table 1. Since time series are often being recorded in different scales, despite the earlier normalization using system size, another round of normalization is in general recommended. After normalization, the series should have zero mean and unit variance. Two features, namely, the mean and variance, can thus be dropped.

The 13 time series features are computed for each of the 405 normalized PV power time series. Using the earlier notation, the data matrix $\boldsymbol{X}$ has a dimension of 405×13. The total process time for computing the generic time series features is 257 s on a late 2013 MacBook Pro. After running PCA, the data points are projected onto the two-dimensional feature space as shown in Figure 3; the numbers in the figure index their corresponding systems. We note that our biplot presentation follows Gabriel (1971), where observations are scaled up by $\sqrt{n}$ and variables scaled down by $\sqrt{n}$. After the projection, two linearly separable clusters can immediately be seen. At this stage, any sensible 2-dimensional clustering algorithm could be used to identify the system types. For our choice, k-means clustering with 2 centers and 25 random initialization of centers is used. Multiple initializations are used because this clustering approach can be sensitive to the initial selection

Table 1: Fifteen non-seasonal time series features used for PV system type identifications. These generic features are adopted from Hyndman et al. (2015).

Feature	Description
Mean	Mean.
Var	Variance.
ACF1	First order of autocorrelation.
Trend	Strength of trend.
Linearity	Strength of linearity.
Curvature	Strength of curvature
Entropy	Spectral entropy.
Lumpiness	Changing variance.
Spikiness	Strength of spikiness.
Lshift	Level shift using rolling window.
Vchange	Variance change.
Fspots	Flat spots using discretization.
Cpoints	The number of crossing points.
KLscore	Kullback-Leibler score.
Change.idx	Index of the maximum KL score.

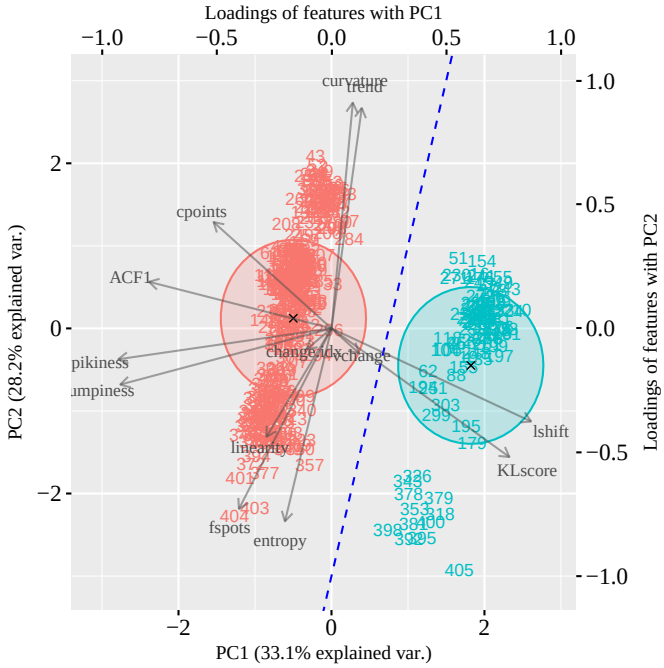

Figure 3: Clustering of PV system types (fixed or tracking) using k-means with principal component analysis. The Indian red cluster shows the PV systems with fixed orientations; the turquoise blue cluster shows the PV systems with trackers. The cluster centers are indicated with black crosses. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of centers. The initialization that leads to the best results[4] is chosen. The resulting clusters are displayed in Figure 3 using Indian red and turquoise, which represent PV systems with fixed orientations and trackers, respectively.

It can be seen from the biplot that some features, namely, vchange and change.idx, are less

[4]Best results refer to the smallest sum of squares of the observations to their assigned cluster centers.

variable than others across all time series; they contribute less in terms of separating the systems. The figure also reveals some opposing features, e.g., curvature and linearity; lshift and cpoints. Time series with high linearity in general expects a low curvature. It is worth to mention that the separation of the two clusters is along the direction of lshift, the level shift using rolling window. As this feature computes the maximum absolute difference between consecutive mean values from a rolling window[5], the tracker systems with more rapid power changes have higher lshift values than those systems with fixed orientations.

A concluding remark of this section is on the time series feature design. As most of the generic time series features used in this application are well distributed in the two-dimensional space, i.e., pointing to various directions, redundancy in the feature space is minimal. A similar distribution of feature vector direction was observed earlier for fast moving consumer goods time series (Yang et al., 2015a). We thus believe that these features designed by Hyndman et al. (2015) are representative and sufficient for many applications.

4. Applications B: Irradiance monitoring network design (application-specific features)

Tobler's first law of geography states that near things are more alike than distant things, and solar irradiance is no exception. In an irradiance monitoring network design problem, we aim to separate a geographical area into strata, so that irradiance at locations in each stratum can be sufficiently represented by data from a single sensor. The immediate task is thus to define a distance measure to gauge the *proximity* between locations, and such measure needs not be geographical distance. Once the proximity is defined, the stratification of the geographical area can be performed using clustering algorithms in the proximity space, or feature space if PCA is considered.

In alignment with the present work is the study carried out by Zagouras et al. (2013), in which the PCA plus k-means framework was used on satellite-derived irradiance maps. In their approach, the satellite-derived irradiance was first converted to a so-called "cloud modification factor (CMF)", which is essentially just clear sky index. Two years (730 days) of daily average values of CMF at 28800 pixels over the greater area of Greece were arranged into a data matrix. After PCA, the 730 initial dimensions were reduced to 102 eigenvectors that were thought to preserve sufficient initial variance. The k-means algorithm was subsequently used to cluster the dimension-reduced dataset into 22 clusters; the optimal number is determined based on two validation indices. In Zagouras et al. (2013)'s case, proximity is the Euclidean distance in the feature space derived by performing PCA.

Although the work is perhaps the first in using PCA to design irradiance monitoring network, it has two drawbacks. The first drawback, as mentioned by Yang and Reindl (2015), is its inflexible design criterion. Since the data matrix is formed by daily CMF values, the optimality of the design is solely based on the geometrical structure of the data. On this point, a method which would consider multiple criteria is desired. Secondly, the dimension-reduced dataset carries little physical meaning, and the feature space cannot be interpreted graphically (102 is still a big number to be plotted). Our present approach could effectively improve on these two drawbacks.

[5]This is the precise reason that we include the code segment in the paper. The computational details can be well understood from the code.

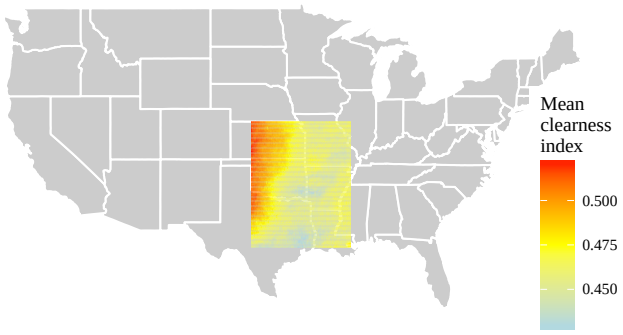

Figure 4: SUNY data used for the network design covers a 10° by 10° square (10000 pixels) over some states in Southern US. The heatmap shows the mean hourly clearness index in 2004.

4.1. SUNY data

The State University of New York (SUNY) gridded satellite-derived irradiance data is used in this application. The full dataset contains hourly estimates (using the Perez et al., 2002, model) of global, diffuse and direct irradiance over a 10 km (about 0.1° latitude and longitude) grid for all states in the United States, except for Alaska where satellite cannot resolve cloud cover information, for 1998 to 2005. A copy of the dataset can be obtained freely at `ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar`.

To demonstrate our algorithm, using a spatio-temporal subset of the full dataset is sufficient. The partial dataset of our choice contains estimates from the year 2004 and covers a 10° by 10° square (10000 pixels) over some states in Southern US. A commonly adopted irradiance data transformation technique, namely, converting global horizontal irradiance to clearness index, k_t, is then applied to the partial dataset for diurnal trend removal. The geographical coverage of the partial dataset and the heatmap of the mean hourly clearness index over that region are shown in Figure 4.

4.2. Choice of features

Recall that we are interested in constructing a proximity so that the similarity in irradiance at different locations can be quantified. More specifically, if each irradiance time series can be represented by a set of characteristics, the proximity will be the Euclidean distance in the (re-expressed) feature space. The most intuitive features under this consideration are perhaps the geographical locations, namely, the latitude and longitude of each pixel. By including the locations as features, the Tobler's first law is reinforced numerically during the analysis.

Besides the geographical locations, the next best approach to generate characteristics is using statistics. Watanabe et al. (2016) used sample mean, variance and entropy to evaluate the variation in solar irradiance. Woyte et al. (2007) used a wavelet-based localized spectral analysis to identify and classify the fluctuations in time series of the instantaneous clearness index. One could give plenty of examples of such statistics for characterizing irradiance. Our network design approach is thus flexible in terms of choice of features.

Generally speaking, a random variable can be well characterized by its distribution function (Kobayashi et al., 2011). There are various ways to describe a distribution function. For instance, we can use quantiles to describe an empirical distribution function, and use parameters to describe a parametric distribution function. In the monitoring network design problem, it is found that some descriptive statistics are more useful than others. After some exploratory analyses, see Appendix B, it is concluded that two features, namely, the lag 12 autocorrelation of the clearness index time series and the highest *location* parameter in the fitted skew-normal mixture distribution[6], are most informative.

Besides statistical features, features with physical and engineering implications can also be considered. For example, it is well-known that the optimal orientation of a PV system, in terms of maximizing its annual yield, is subjected to not only the Sun path, but also the intricate geographical and climatic conditions. Many works have shown that by deviating from the conventional PV placement strategy (latitude-level tilt and equator-facing), the annual energy yield of a PV system can be improved significantly (Smith et al., 2016; Lave et al., 2015; Khoo et al., 2014). More details on optimizing PV orientation are provided in Appendix C and the references therein cited. Suppose one of the tasks of the designed monitoring network is to help monitor the PV performance in its proximity, features such as the optimal tilt and azimuth angles are useful. These two features, together with the earlier features, are arranged in Table 2; four feature maps are plotted in Figure 5. It is observed that all four features contain strong spatial structure (see discussion below), which is in favor of the linearity assumption in PCA.

Table 2: Six application-specific time series features used for irradiance monitoring network design. See Appendix B and Appendix C for more details.

Feature	Description
Lat	Latitude.
Lon	Longitude.
ACF12	Lag 12 autocorrelation.
μ_1	The highest location in the fitted skew-normal mixture distribution.
Opt.Tilt	Optimal PV tilt angle that maximizes annual yield.
Opt.Azimuth	Optimal PV azimuth angle that maximizes annual yield.

4.3. Clustering results

The data matrix $\boldsymbol{X}$ in this application has a dimension of 10000×6. Once the data matrix is prepared, PCA is performed and the corresponding biplot is shown in the left panel of Figure 6. Unlike the previous application, it is observed that the clusters are not linearly separable in this case. It is therefore necessary to choose a "k" during clustering. As mentioned earlier, Zagouras et al. (2013) computed two indices, namely, the DB index (Davies and Bouldin, 1979) and CH index (Caliński and Harabasz, 1974). As both indices decrease with the number of clusters, the elbow method (Thorndike, 1953) was then used to identify the optimal value of k. Besides these two

[6]It is known that the distribution of clearness index is bimodal. Mixture distributions are often used to model distributions with two peaks. The highest location parameter in the mixture distribution is essentially gauging the position of the second peak in the bimodal distribution.

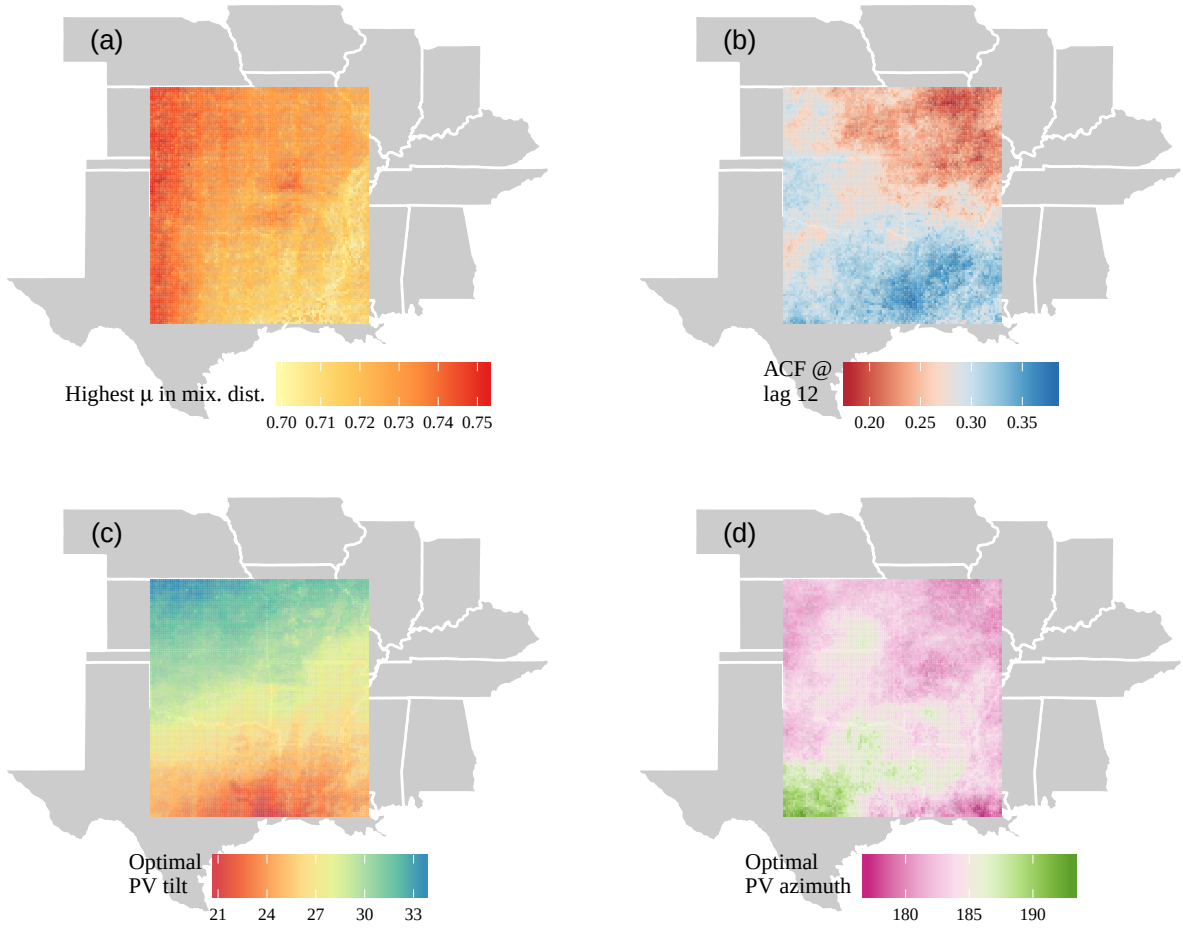

Figure 5: Maps of four PCA features extracted from k_t time series. (a) The highest mean value (location parameter) from the mixture distribution fitting; (b) the autocorrelation at lag 12; (c, d) optimal PV tilt and azimuth angles that maximize the annual yield of a flat surface collector.

indices, the elbow method can be applied to many other evaluation metrics, such as the percentage of variance explained (Goutte et al., 1999) and Silhouette index (Rousseeuw, 1987). However, it is not our immediate interest to advise on the "most appropriate" validation index for irradiance monitoring network design in this work. Instead, a fixed number of clusters, 10, is adopted. This fixed number can be thought of in a practical context as the number of sensors that an installation budget allows. Based on the setting of $k = 10$, the final cluster map is shown on the right panel of Figure 6.

The clustering results show a series of important findings. Firstly, latitude and longitude appear to be the major features defining the first and second principal components, respectively. This indicates that the geographical location contributes the most to the overall variation in the time series feature space, i.e., a major deciding factor in our network design setup. Secondly, the small angle between latitude and the optimal PV tilt suggests that the optimal PV tilt depends largely on the site's latitude (the two variables are highly correlated), but not on longitude; this

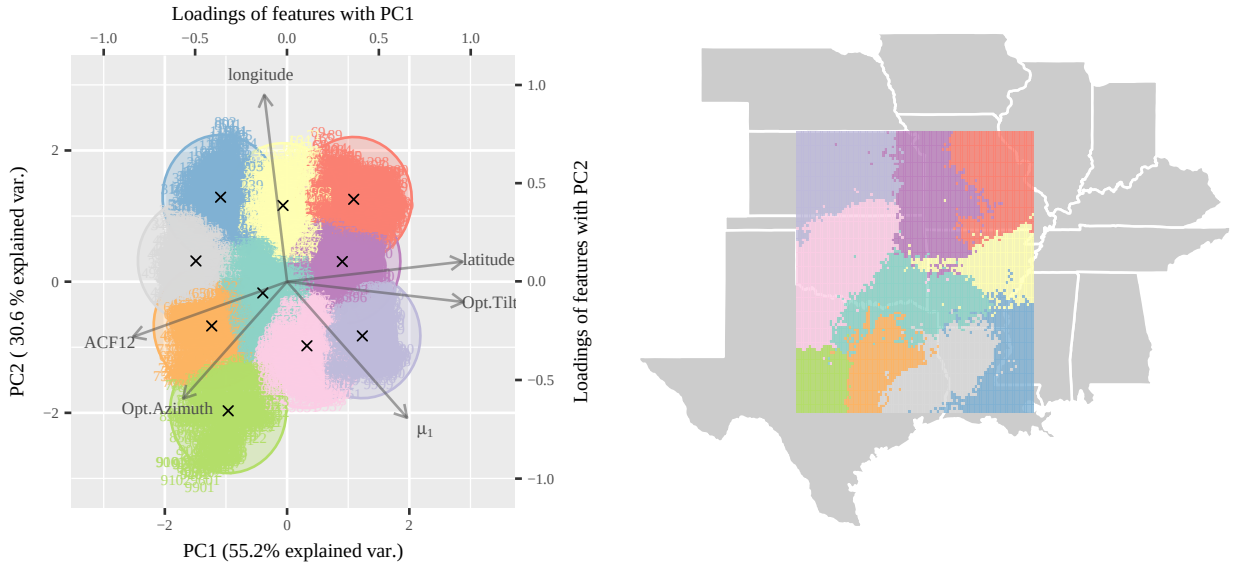

Figure 6: (Left) Biplot of the satellite-derived irradiance data; a total of 6 features are considered during PCA. (Right) The k-means clustering results. Pixels with same colors have similar properties, and thus can be approximated by a single sensor.

knowledge is known *a priori*. Another important observation is that the clustering results shown in Figure 6 agree well with the features maps. For example, the chartreuse cluster along the vector representing optimal PV azimuth can be related to the observations shown in the bottom left corner of Figure 5 (d), namely, the lime green patch; the pink and light steel blue clusters along the vector μ_1 can be linked to the high μ_1 values depicted in the left side of Figure 5 (a). We conclude that the geographical shapes of the final clusters are influenced by the selected features. This linkage between the feature maps and clustering results gives flexibility and practical advantages to our network design approach. The design framework can be applied to any features that are thought appropriate, and the results are readily interpretable.

5. Applications C: PV string fault detection (exploratory features)

In the era of data, solar engineers are often required to study PV data for tasks such as PV degradation estimation, monitoring quality control and string fault identification. Many traditional data processing techniques, which examine a few time series of a few parameters at a time, can be inadequate when dealing with large amount of data. For instance, identifying a faulty string from a utility-scale PV plant with thousands of strings is difficult; this section considers such scenarios. In particular, we analyze data from a PV plant located in Senftenberg, Germany. The monthly files contain 15 min information including string-level current, voltage, main switch status, surge protector status and combiner box temperature.

In the literature, many PV fault detection methods involve comparing a signal (power, voltage or current) to its theoretical expectation (e.g., Dhimish and Holmes, 2016; Chine et al., 2016; Platon et al., 2015). Such methods mostly likely would require meteorological measurements such as irradiance and ambient temperature, which may not be always available. An alternative

to these methods is to use peers to identify anomalous strings with statistical outlier detection rules (e.g., Zhao et al., 2014, 2013). The statistical outlier detection is usually performed in a univariate setting, i.e., based on a single variable such as current in a string. In other words, this type of method is somewhat limited to detecting faults that can be associated to a single variable. However, as Chine et al. (2016) summarized in Table 2 of their work, deviations observed in parameters such as string output current, voltage and short circuit current can be linked back to more than one type of fault. A statistical outlier detection method that can operate in multivariate settings is thus desired. On this point, the approach discussed in this paper is suitable.

Similar to the previous applications, a set of features is first defined based on domain knowledge or experts' view. After the PCA, the data are again represented in the feature space in the form of a biplot. Standard two-dimensional anomaly detection algorithms such as highest density regions and α-null can then be applied to identify outliers in the space formed by first and second PCs. This approach was first discussed by Hyndman et al. (2015). In this way, the faulty string could be detected based on its "overall" anomalousness. Detailed fault identification and root cause analysis could be carried out subsequently.

5.1. Feature selection

The power plant in Senftenberg is connected to PVGuard WebPortal[7], a online monitoring platform that gained popularity through the past decade. The web portal itself is equipped with the capability of generating fault alarms up to string level. We examine some frequently reported faults to select our time series features; these faults include:

1. Inverter power drop-out, current Power: 0 kW;
2. DC Current: String x differs from the other strings by y%;
3. Combiner box status event: Voltage Deviation;
4. Combiner box status event: Power Drop;
5. Inverter fault state: IGBT Switch Fault.

It can be concluded that most faults can be detected by observing the output current and voltage at difference levels, i.e., inverter, combiner box and string levels. It is also noted that if I–V curves are available, more faults can be identified (Chine et al., 2016). Unfortunately, our dataset does not contain I–V curve information. To that end, the four features considered for this part of the work are listed in Table 3. Each feature can be evaluated over a period of T, which is taken as one week in this case study. A total of 14 ($= 4 \times 4 - 2$) features are generated using the four weeks of data from 2017 March. Two weeks in that month contain no missing data, feature F1 from these two weeks are thus excluded from the data matrix $\boldsymbol{X}$.

5.2. Outlier detection

After performing PCA, the biplot of this application is depicted in Figure 7. A total 1391 points are indexed in the plot, representing the 1391 strings in the Senftenberg system. Some anomalous

[7] `http://www.skytron-energy.com/en/system/supervision-software/pv-scada-software-webportal/`

Table 3: Four exploratory time series features used for PV string fault detection.

Feature	Description
F1	Number of missing data points.
F2	Number of data points that record a negative current,.
F3	Mean current in a string.
F4	Mean voltage across a string.

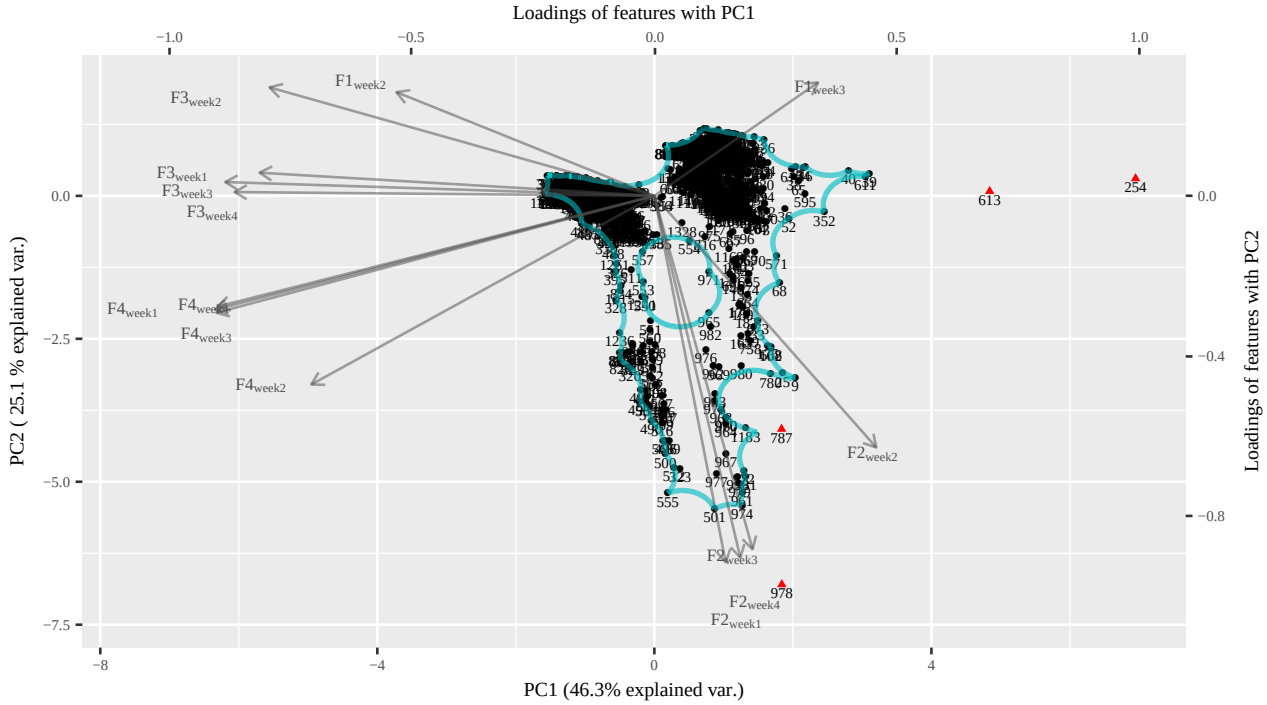

Figure 7: Biplot for the anomalous PV string detection application. Four features, names, F1 to F4, are extracted from four weeks of data from 2017 March. The α-hull algorithm, with $\alpha = 0.5$, is used for outlier detection. The outliers are marked with Indian red triangles; the main hull is contoured with turquoise blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

strings, such as string numbers 254 and 613, can be spotted immediately without any detection algorithm. However, there are many definitions of an outlier in the literature (e.g., Barnett and Lewis, 1994; Hawkins, 1980). There is not any universal rule that governs what should be called an outlier. Under such considerations, we prefer sequential outlier detection methods (identify one or few outliers at a time) over the single-step methods (identifying all outliers at once), which in general grants more control over the outlier detection and provides a rank of anomalousness in the outliers detected.

We consider the α-hull algorithm (see Appendix D for more details). When the control parameter, namely, α, decreases from ∞ to 0, a continuous transform from the convex hull to isolated points (singletons) is observed. Based on this property, the α-hull algorithm can be used as a sequential two-dimensional outlier detection algorithm. An example hull with $\alpha = 0.5$ is shown in Figure 7. Four singletons have been isolated at this stage.

It can be seen from Figure 7 that features F3 computed from different weeks are well correlated with each other, i.e., pointing to similar directions. Similar observations can be found for F2 and F4, but not for F1. Recall the features defined in Table 3, high correlation in features from different weeks reveal that various faults persist through the month. String numbers 254 and 613 locate along the opposing direction of F3, indicating the mean current values in these strings are low. String numbers 978 and 787 locate along the direction of F2, indicating that there are some negative current measurements in these strings. At this stage, four outliers have been identified; a quick plot of data confirms the findings (we omit the plot). The proposed detection method is shown to be able to narrow down the unusual strings, so that detailed investigation can be performed subsequently.

6. Concluding remarks

An analytics method for handling big time series data is discussed. In this method, each time series is reduced to a set of features, either generic or application-specific. The reduced feature space facilitates visualization and analyses. Three solar engineering applications are considered to demonstrate the idea. Traditional approaches to these applications consider data points as individuals; the present approach that considers time series as entities is thus novel in terms of data handling. Principal component analysis and biplot are the main tools in all three applications. Biplots make the results of PCA interpretable. By examining the biplots, geometrical relationships among the features and original time series can be established, which leads to insights that are otherwise unobservable using traditional methods.

The analytics method is flexible in terms of feature design and can be applied to a variety of other applications. However, common to all dimension reduction strategies, extracting time series features may result in information loss. One should be cautious when replacing the traditional methods with the present method. We advise readers to use the proposed method as an exploratory tool rather than solely relying on the method itself.

Conflict of Intrest

Authors declare no conflict of interest.

Acknowledgment

This work is partially supported under the A*STAR TSRP fund 1424200021 and Antuit–SIMTech Supply Chain Analytics Lab.

Many thanks go to J. Y. Li, who graduated as a Master of International Relations from the University of Melbourne, for pointing out the ambiguities in the earlier version of the paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at *doi place holder*. We will include the supplementary data should the paper be accepted.

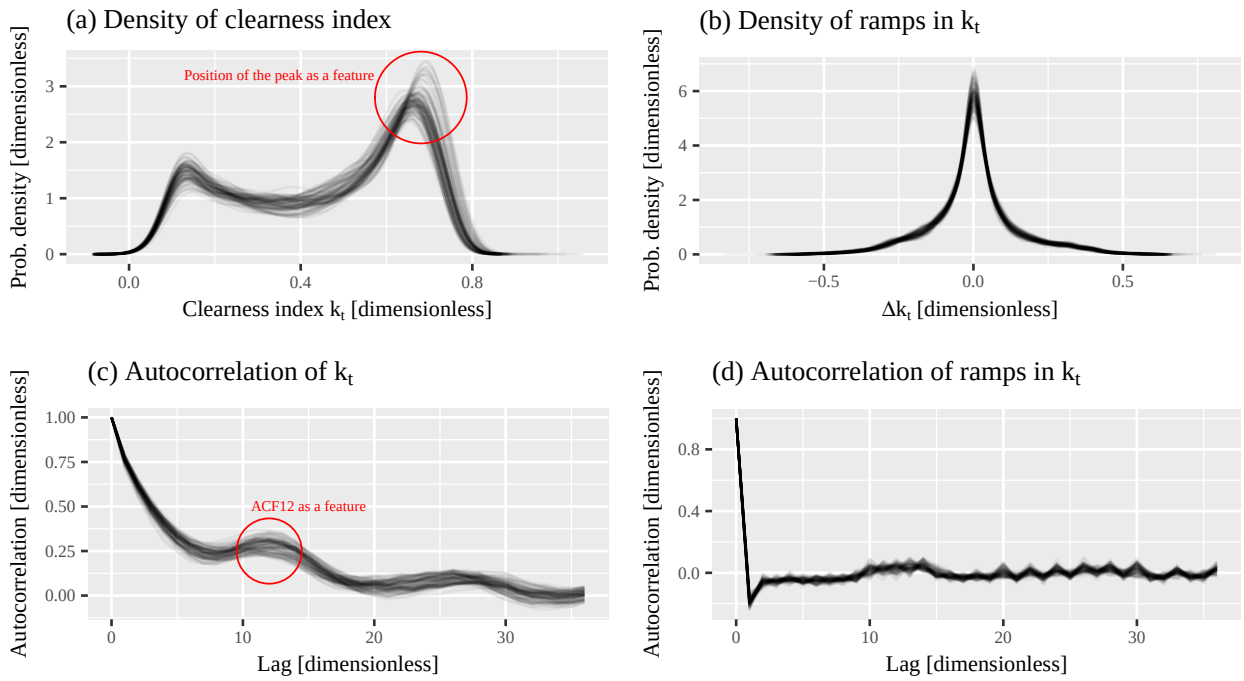

Figure B.8: Statistical characteristics of 100 randomly selected clearness index time series. Subplots (a, b) show the probability density functions of k_t and Δk_t, respectively; subplots (c, d) show the autocorrelation functions of k_t and Δk_t, respectively. The pdf of k_t is shown to be bimodal; the pdfs of various time series differ mostly at their second peaks.

Appendix B. Exploratory analysis on distributions of clearness index

It is considered appropriate that an irradiance (or clearness index) time series can be characterized through statistical distributions. Where a parametric distribution is concerned, its parameters describe the distribution (e.g., μ and σ describe a normal distribution). When the distribution is taken as an empirical one, statistics such as quantiles can be used to summarize the distribution (e.g., 25, 50, 75 percentiles). In either case, the original time series is reduced to a few numbers (parameters and/or quantiles), which can be taken as features for our irradiance monitoring network design application. Besides the distribution itself, other statistics can also be employed. On this point, Hansen et al. (2010) suggested three statistics, namely,

1. Statistical distribution of irradiance (or clearness index);
2. Statistical distribution of ramps of irradiance (or clearness index);
3. Autocorrelations functions of clearness index and of ramps in clearness index.

This appendix aims to explore these statistics visually and select the most meaningful features, so that satellite pixels with similar time series features could be subsequently grouped together.

Figure B.8 (a) shows the kernel density estimates (KDE) of distributions of 100 randomly selected k_t time series, while subplot (b) shows the KDE for the corresponding Δk_t distribution. Figure B.8 (c) and (d) show the autocorrelation functions of those selected k_t and Δk_t time series, respectively. As our goal is to use features to describe each time series for clustering purpose, it

is more amenable if the features are descriptive and distinctive. To that end, the information contained in density and autocorrelation function of Δk_t time series can be dropped immediately due to the high similarities among various time series. On the other hand, as annotated in Figure B.8 (c), it is observed that the spread of autocorrelation functions of k_t is most significant around lag 12. We therefore select ACF12 as one of our features.

It is also noticeable from Figure B.8 (a) that the distributions of k_t time series are most distinguishable around $k_t = 0.7$, i.e., the higher peaks in the KDE. A parameter which can describe the position of this peak is thus needed. More particularly, given the bimodal distribution, we are looking for the "center" of the largest component of some finite mixtures of distributions. A mixture distribution has density of the form

$$g(z; \mathbf{\Theta}) = \sum_{i=1}^{n} p_i f(z; \boldsymbol{\theta}_i), \tag{B.1}$$

where $p_i \geq 0$, with $\sum_{i=1}^{n} p_i = 1$, are the mixing weights; $f(\cdot; \boldsymbol{\theta}_i)$ is density of the i-th component in the mixture, parameterized by $\boldsymbol{\theta}_i$; and $\mathbf{\Theta} = (p_1, \cdots, p_n, \boldsymbol{\theta}_1^\top, \cdots, \boldsymbol{\theta}_n^\top)^\top$. In describing the irradiance distribution, the value for n, i.e., the number of components, it is often taken as 2 or 3.

In fact, the statistical distribution of solar irradiance, or clearness index, has been well-studied for data over a wide range of temporal granularities (Voskrebenzev et al., 2015; Hollands and Suehrcke, 2013; Jurado et al., 1995; Saunier et al., 1987; Hollands and Huget, 1983). Most of these works confirm that the distribution of clearness index follows a bimodal distribution, originated from the clear and cloudy radiation states. Among various modeling approaches, mixtures of normal distributions are most commonly used. For examples, Jurado et al. (1995); Hollands and Suehrcke (2013) use two- and three-component normal mixtures, respectively. Figure B.9 shows the fittings of two- and three-component normal mixture distributions to four randomly selected clearness index time series. The results are however unsatisfactory, due to the skews observed at both high and low k_t values. Furthermore, the bimodal distribution is found having a positive skewness at low k_t and a negative skewness at high k_t values. To model this, we consider the scale mixtures of skew-normal distribution.

A univariate random variable Z has skew-normal distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ, if its density is given by

$$f(z; \boldsymbol{\theta}) = 2\phi\left(z; \mu, \sigma^2\right) \Phi\left(\frac{\lambda(z-\mu)}{\sigma}\right), \tag{B.2}$$

where $\boldsymbol{\theta} = (\mu, \sigma^2, \lambda)^\top$; ϕ is the probability density function of a normal distribution with mean μ and variance σ^2; and Φ is cumulative distribution function of the standard normal distribution. If we assume the mixture distribution can be modeled with three components, its parameter $\mathbf{\Theta} = (p_1, p_2, p_3, \boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_3^\top)^\top$, where $\boldsymbol{\theta_i} = (\mu_i, \sigma_i^2, \lambda_i)^\top$, can be determined by fitting the parametric distribution to observations. In this paper, parameter $\mathbf{\Theta}$ is estimated via the expectation–maximization algorithm, as described in Prates et al. (2013), for each clearness index time series. Four examples of the fitted density are plotted in Figure B.9. It can be seen that the skew-normal mixtures fit the empirical distributions better than two- or three-component normal mixtures. In this way, we obtain our second feature, namely, $\mu^* = \max(\mu_1, \mu_2, \mu_3)$. This feature, together with ACF12, is plotted in Figure 5.

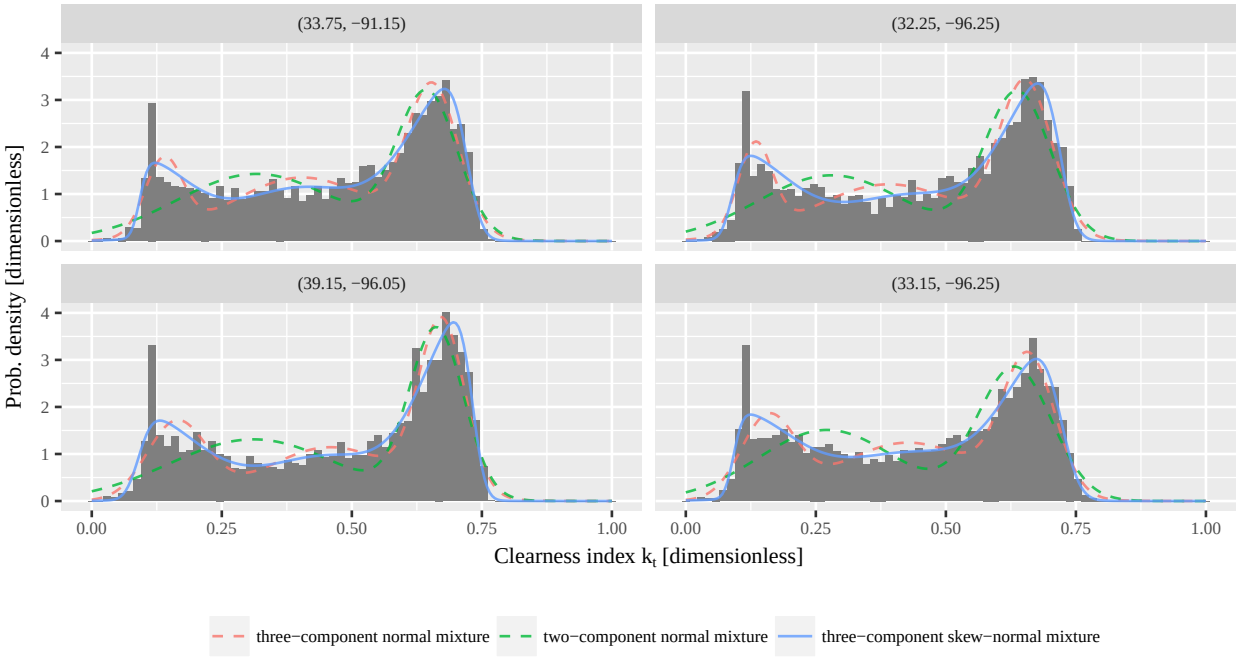

Figure B.9: Distributions of 4 randomly selected clearness index time series. The scale mixture of skew-normal distributions with three components shows best fit.

Appendix C. Determining the optimal orientation for a flat surface solar collector

To maximize the annual yield of a flat surface solar collector, the collector is often placed at a tilt angle equals to the latitude of the site, facing the equator. However, due to the intricate location-specific geographical and climatic conditions, this rule-of-thumb may not be always optimal. For such reasons, simulation is used to optimize the orientation and thus maximize the expected energy output of a solar collector (Smith et al., 2016; Lave et al., 2015; Khoo et al., 2014).

The procedure of the simulation usually involves a so-called "transposition model", which converts a set of irradiance measurements collected on a horizontal surface to the irradiance expected on a tilted surface. For every pair of tilt (α) and azimuth (β) angles, the global tilted irradiance (G_c) can be simulated. In this way, the optimization problem can be written as:

$$\underset{\alpha,\beta}{\operatorname{argmax}} \sum_{t=1}^{T} \widehat{G}_c(t), \tag{C.1}$$

where $\widehat{G}_c(t)$ is the modeled global tilted irradiance.

There are many transposition models available; most of them require the information of global horizontal irradiance and diffuse horizontal irradiance. The transposition models have varying degree of complexity, we refer the readers to the review by Yang (2016) for a detailed comparison of transposition models. However, for our current purpose, we adopt the simplest model, namely, the isotropic model (Liu and Jordan, 1961), in our simulation. The maximization problem is solved

using the general-purpose optimization routine (function `optim`) in R (R Core Team, 2016). The detailed procedure of our simulation and the computational issues are discussed in Yang et al. (2016). After the pair optimal tile and azimuth angles is found for each pixel, the two features (α and β) are plotted in Figure 5.

Appendix D. The α-hull algorithm

The α-convex hull, or simply α-hull, is a generalized convex hull[8] that is studied in different fields of research, primarily in computational geometry. Informally, a hull can be considered as a geometrical structure that serves to characterize the shape of a set. Unlike the convex hull, the α-hull is able to reconstruct non-convex sets. The mathematical definition of α-hull is given by Pateiro-Lopez and Rodriguez-Casal (2016): if we denote an open ball with center x and radius α with $\mathring{B}(x, \alpha)$, a set $A \subset \mathbb{R}^d$ is said to be α-convex, for $\alpha > 0$, if

$$A = C_\alpha(A) = \bigcap_{\{\mathring{B}(x,\alpha):\mathring{B}(x,\alpha)\cap A=\emptyset\}} \left(\mathring{B}(x, \alpha)\right)^c, \tag{D.1}$$

where $C_\alpha(A)$ is called the α-hull of A. We can immediately see that the shape of the hull is influenced by the ball radius α.

Starting from some big enough initial α value which returns a hull that bounds all points, a series of α-hulls are computed in the space formed by first two PCs with decreasing α. A point is identified as an outlier when it becomes a singleton. With the decreasing α, the outliers identified in each step are naturally assigned with a rank; the highest rank corresponds to the α value that leads to the first singleton. The implementation of the α-hull invoves the Delaunay triangulation, Voronoi diagram and α-shape. We use the R package `alphahull` (Pateiro-Lopez and Rodriguez-Casal, 2016) to compute α-hulls for our application.

Appendix E. Several other potential applications of the analytics algorithm

Three applications are demonstrated in the main text. We outline several other potential applications that can benefit from the PCA-based algorithm.

Appendix E.1. Forecasting with sensor network

Irradiance forecasting is an important aspect towards integrating variable solar energy into the electricity grid reliably and efficiently. There is a rich literature in solar forecasting; a good overview on this topic is found in Inman et al. (2013). With the advent of sensor network technology, spatio-temporal forecasting methods became popular. As solar irradiance is a spatio-temporal process, there is very little reason to use univariate methods when data from multiple sensors within a certain geographical proximity are available (see forecast comparisons made by Aryaputera et al., 2015b; Yang et al., 2015c).

[8]The convex hull of a set of points, X, in an affine space is the smallest convex set that contains X. In such a space, the convex set is a subset of that space that is closed under convex combination.

The variability in irradiance is primarily caused by moving clouds. Suppose a regression approach is used for prediction, i.e., predicting the irradiance at a focal station using lagged time series measured at neighboring stations, the direction and speed of these moving clouds have strong implication on predictor selection. In a previous work, the lasso regression was used to make such variable selection (Yang et al., 2015c). However, it was found that even with only 17 stations, the predictor pool could be large, owing to the fact that the number of predictors multiplies when lagged versions of a variable are considered. For example, if time series up to lag-10 from each of the 17 stations are used, the total number of predictors is 170, including the autocorrelated series. On this point, we can consider the approach discussed in this work and pre-select some predictors based on their similarities to the predictand. We refer the readers to a parallel work by Kim and Swanson (2016) for further reading.

Appendix E.2. Forecasting with satellite-derived irradiance

As an alternative to an irradiance sensor, satellite images provide estimates of ground level irradiance. When ground-based measurements are not available, after some bias corrections, satellite-derived irradiance data can be used for forecasting with a horizon up to a few days. One of the options to perform such forecasting is to used time series models such as the seasonal autoregressive integrated moving average model (SARIMA) and exponential smoothing (ETS) model, as seen in Aryaputera et al. (2015a); Yang et al. (2015b, 2012).

These statistical forecasting modals require parameter estimation, e.g., process orders in an SARIMA model; the parameters are periodically updated when rolling forecasts are used. When irradiance values at a large number of locations need to be forecast, training and iteration can be time consuming. If we cluster the time series based on features such as autocorrelation and partial autocorrelation, process orders in a cluster can be assumed similar. This could reduce the search space for 'optimal' process order based on information criteria, which are often used in automated forecasting.

Appendix E.3. PV performance evaluation

Projects like "solar city", in Ōta, Tokyo, prove allure of Sun's energy in Japan. As three quarters of homes in Ōta city's Pal town neighborhood are equipped with rooftop PV systems that are free for their residents, centralized monitoring is important for the investors to ensure the performance of these PV systems. Similar to the application discussed in Section 5, anomalous PV systems can be identified using the proposed method with a set of properly designed evaluation indices, such as performance ratio and yield.

References

Aryaputera, A.W., Yang, D., Walsh, W.M., 2015a. Day-ahead solar irradiance forecasting in a tropical environment. Journal of Solar Energy Engineering 137, 051009 – 051009. doi:http://dx.doi.org/10.1115/1.4030231.

Aryaputera, A.W., Yang, D., Zhao, L., Walsh, W.M., 2015b. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. Solar Energy 122, 1266 – 1278. doi:http://doi.org/10.1016/j.solener.2015.10.023.

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data. J. Wiley & Sons, London.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics 3, 1 – 27. doi:http://dx.doi.org/10.1080/03610927408827101.

Chine, W., Mellit, A., Lughi, V., Malek, A., Sulligoi, G., Pavan, A.M., 2016. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. Renewable Energy 90, 501 – 512. doi:http://dx.doi.org/10.1016/j.renene.2016.01.036.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1, 224 – 227. doi:http://dx.doi.org/10.1109/TPAMI.1979.4766909.

Dhimish, M., Holmes, V., 2016. Fault detection algorithm for grid-connected photovoltaic plants. Solar Energy 137, 236 – 245. doi:http://dx.doi.org/10.1016/j.solener.2016.08.021.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453 – 467. doi:http://dx.doi.org/10.1093/biomet/58.3.453.

GE Energy, 2010. Western Wind and Solar Integration Study. Subcontract report for National Renewable Energy Laboratory NREL/SR-550-47434. GE Energy Management. Schenectady, New York. URL: http://www.nrel.gov/docs/fy10osti/47434.pdf.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F.Å., Hansen, L.K., 1999. On clustering fMRI time series. NeuroImage 9, 298 – 310. doi:http://dx.doi.org/10.1006/nimg.1998.0391.

Guerriero, P., Napoli, F.D., Vallone, G., d'Alessandro, V., Daliento, S., 2016. Monitoring and diagnostics of PV plants by a wireless self-powered sensor for individual panels. IEEE Journal of Photovoltaics 6, 286 – 294. doi:http://dx.doi.org/10.1109/JPHOTOV.2015.2484961.

Hansen, C.W., Stein, J.S., Ellis, A., 2010. Statistical Criteria for Characterizing Irradiance Time Series. Technical Report SAND2010-7314. Sandia National Laboratories. Albuquerque, New Mexico. URL: http://energy.sandia.gov/wp-content/gallery/uploads/107314.pdf.

Hawkins, D.M., 1980. Identification of Outliers. Chapman and Hall, London.

Hollands, K., Huget, R., 1983. A probability density function for the clearness index, with applications. Solar Energy 30, 195 – 209. doi:http://dx.doi.org/10.1016/0038-092X(83)90149-4.

Hollands, K.G.T., Suehrcke, H., 2013. A three-state model for the probability distribution of instantaneous solar radiation, with applications. Solar Energy 96, 103 – 112. doi:http://dx.doi.org/10.1016/j.solener.2013.07.007.

Hummon, M., Ibanez, E., Brinkman, G., Lew, D., 2012. Sub-hour solar data for power system modeling from static spatial variability analysis, in: 2nd International Workshop on Integration of Solar Power in Power Systems, Lisbon, Portugal.

Hyndman, R.J., Wang, E., Laptev, N., 2015. Large-scale unusual time series detection, in: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1616 – 1619. doi:http://dx.doi.org/10.1109/ICDMW.2015.104.

Inman, R.H., Pedro, H.T., Coimbra, C.F., 2013. Solar forecasting methods for renewable energy integration. Progress in Energy and Combustion Science 39, 535 – 576. doi:http://doi.org/10.1016/j.pecs.2013.06.002.

Jolliffe, I., 2002. Principal component analysis. Springer Verlag, New York.

Jurado, M., Caridad, J., Ruiz, V., 1995. Statistical distribution of the clearness index with radiation data integrated over five minute intervals. Solar Energy 55, 469 – 473. doi:http://dx.doi.org/10.1016/0038-092X(95)00067-2.

Khoo, Y.S., Nobre, A., Malhotra, R., Yang, D., Rüther, R., Reindl, T., Aberle, A.G., 2014. Optimal orientation and tilt angle for maximizing in-plane solar irradiation for PV applications in Singapore. IEEE Journal of Photovoltaics 4, 647 – 653. doi:http://dx.doi.org/10.1109/JPHOTOV.2013.2292743.

Kim, H.H., Swanson, N.R., 2016. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. International Journal of Forecasting (In Press). doi:http://doi.org/10.1016/j.ijforecast.2016.02.012.

Kobayashi, H., Mark, B.L., Turin, W., 2011. Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance. Cambridge University Press. doi:http://dx.doi.org/10.1017/CBO9780511977770.

Lave, M., Hayes, W., Pohl, A., Hansen, C.W., 2015. Evaluation of global horizontal irradiance to plane-of-array irradiance models at locations across the United States. IEEE Journal of Photovoltaics 5, 597 – 606. doi:http://dx.doi.org/10.1109/JPHOTOV.2015.2392938.

Lew, D., Brinkman, G., Ibanez, E., Florita, A., Heaney, M., Hodge, B.M., Hummon, M., Stark, G., King, J., Lefton,

S.A., Kumar, N., Agen, D., Jordan, G., Venkataraman, S., 2013. The Western Wind and Solar Integration Study Phase 2. Technical Report NREL/TP-5500-55588. National Renewable Energy Laboratory. Golden, Colorado. URL: http://www.nrel.gov/docs/fy13osti/55588.pdf.

Liu, B.Y.H., Jordan, R.C., 1961. Daily insolation on surfaces tilted towards the equator. ASHRAE Transactions 67, 526 – 541.

Miller, N.W., Shao, M., Pajic, S., D'Aquila, R., 2014. Western Wind and Solar Integration Study Phase 3 - Frequency Response and Transient Stability. Subcontract report for National Renewable Energy Laboratory NREL/SR-5D00-62906. GE Energy Management. Schenectady, New York. URL: http://www.nrel.gov/docs/fy15osti/62906.pdf.

Mousazadeh, H., Keyhani, A., Javadi, A., Mobli, H., Abrinia, K., Sharifi, A., 2009. A review of principle and sun-tracking methods for maximizing solar systems output. Renewable and Sustainable Energy Reviews 13, 1800 – 1818. doi:http://dx.doi.org/10.1016/j.rser.2009.01.022.

Nann, S., 1990. Potentials for tracking photovoltaic systems and V-troughs in moderate climates. Solar Energy 45, 385 – 393. doi:http://dx.doi.org/10.1016/0038-092X(90)90160-E.

Nikitidou, E., Kazantzidis, A., Tzoumanikas, P., Salamalikis, V., Bais, A., 2015. Retrieval of surface solar irradiance, based on satellite-derived cloud information, in Greece. Energy 90, Part 1, 776 – 783. doi:http://doi.org/10.1016/j.energy.2015.07.103.

Pateiro-Lopez, B., Rodriguez-Casal, A., 2016. alphahull: Generalization of the Convex Hull of a Sample of Points in the Plane. URL: https://CRAN.R-project.org/package=alphahull. r package version 2.1.

Perez, R., Ineichen, P., Moore, K., Kmiecik, M., Chain, C., George, R., Vignola, F., 2002. A new operational model for satellite-derived irradiances: description and validation. Solar Energy 73, 307 – 317. doi:http://dx.doi.org/10.1016/S0038-092X(02)00122-6.

Platon, R., Martel, J., Woodruff, N., Chau, T.Y., 2015. Online fault detection in PV systems. IEEE Transactions on Sustainable Energy 6, 1200 – 1207. doi:http://dx.doi.org/10.1109/TSTE.2015.2421447.

Prates, M.O., Cabral, C.R.B., Lachos, V.H., 2013. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. Journal of Statistical Software 54, 1 – 20. URL: http://www.jstatsoft.org/v54/i12/.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53 – 65. doi:http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Saunier, G., Reddy, T., Kumar, S., 1987. A monthly probability distribution function of daily global irradiation values appropriate for both tropical and temperate locations. Solar Energy 38, 169 – 177. doi:http://dx.doi.org/10.1016/0038-092X(87)90015-6.

Shlens, J., 2003. A tutorial on principal component analysis. https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf. Accessed: 2017-04-17.

Smith, C.J., Forster, P.M., Crook, R., 2016. An all-sky radiative transfer method to predict optimal tilt and azimuth angle of a solar collector. Solar Energy 123, 88 – 101. doi:http://dx.doi.org/10.1016/j.solener.2015.11.013.

Thorndike, R.L., 1953. Who belongs in the family? Psychometrika 18, 267 – 276. doi:http://dx.doi.org/10.1007/BF02289263.

Voskrebenzev, A., Riechelmann, S., Bais, A., Slaper, H., Seckmeyer, G., 2015. Estimating probability distributions of solar irradiance. Theoretical and Applied Climatology 119, 465 – 479. doi:http://dx.doi.org/10.1007/s00704-014-1189-9.

Watanabe, T., Takamatsu, T., Nakajima, T.Y., 2016. Evaluation of variation in surface solar irradiance and clustering of observation stations in Japan. Journal of Applied Meteorology and Climatology 55, 2165 – 2180. doi:http://dx.doi.org/10.1175/JAMC-D-15-0227.1.

Wilcox, S., Marion, W., 2008. Users Manual for TMY3 Data Sets. Technical Report NREL/TP-581-43156. National Renewable Energy Laboratory. Golden, Colorado. URL: http://www.nrel.gov/docs/fy08osti/43156.pdf.

Woyte, A., Belmans, R., Nijs, J., 2007. Fluctuations in instantaneous clearness index: Analysis and statistics. Solar Energy 81, 195 – 206. doi:http://dx.doi.org/10.1016/j.solener.2006.03.001.

Wu, X., Zhu, X., Wu, G.Q., Ding, W., 2014. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering 26, 97 – 107. doi:http://dx.doi.org/10.1109/TKDE.2013.109.

Yang, D., 2016. Solar radiation on inclined surfaces: Corrections and benchmarks. Solar Energy 136, 288 – 302. doi:http://dx.doi.org/10.1016/j.solener.2016.06.062.

Yang, D., Goh, G.S.W., Jiang, S., Zhang, A.N., 2016. Spatial data dimension reduction using quadtree: A case study on satellite-derived solar radiation, in: 2016 IEEE International Conference on Big Data (Big Data), pp. 3807 – 3812. doi:http://dx.doi.org/10.1109/BigData.2016.7841052.

Yang, D., Goh, G.S.W., Xu, C., Zhang, A.N., Akcan, O., 2015a. Forecast UPC-level FMCG demand, Part I: Exploratory analysis and visualization, in: 2015 IEEE International Conference on Big Data (Big Data), pp. 2106 – 2112. doi:http://dx.doi.org/10.1109/BigData.2015.7363993.

Yang, D., Jirutitijaroen, P., Walsh, W.M., 2012. Hourly solar irradiance time series forecasting using cloud cover index. Solar Energy 86, 3531 – 3543. doi:http://doi.org/10.1016/j.solener.2012.07.029. Solar Resources.

Yang, D., Quan, H., Disfani, V.R., Liu, L., 2017. Reconciling solar forecasts: Geographical hierarchy. Solar Energy 146, 276 – 286. doi:http://doi.org/10.1016/j.solener.2017.02.010.

Yang, D., Reindl, T., 2015. Solar irradiance monitoring network design using the variance quadtree algorithm. Renewables: Wind, Water, and Solar 2, 1 – 8.

Yang, D., Sharma, V., Ye, Z., Lim, L.I., Zhao, L., Aryaputera, A.W., 2015b. Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. Energy 81, 111 – 119. doi:http://dx.doi.org/10.1016/j.energy.2014.11.082.

Yang, D., Ye, Z., Lim, L.H.I., Dong, Z., 2015c. Very short term irradiance forecasting using the lasso. Solar Energy 114, 314 – 326. doi:http://doi.org/10.1016/j.solener.2015.01.016.

Zagouras, A., Kazantzidis, A., Nikitidou, E., Argiriou, A., 2013. Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. Solar Energy 97, 1 – 11.

Zhao, Y., Balboni, F., Arnaud, T., Mosesian, J., Ball, R., Lehman, B., 2014. Fault experiments in a commercial-scale PV laboratory and fault detection using local outlier factor, in: 2014 IEEE 40th Photovoltaic Specialist Conference (PVSC), pp. 3398 – 3403. doi:http://dx.doi.org/10.1109/PVSC.2014.6925661.

Zhao, Y., Lehman, B., Ball, R., Mosesian, J., de Palma, J.F., 2013. Outlier detection rules for fault detection in solar photovoltaic arrays, in: 2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC), pp. 2913 – 2920. doi:http://dx.doi.org/10.1109/APEC.2013.6520712.