Evaluation of Variation in Surface Solar Irradiance and Clustering of Observation Stations in Japan

Takeshi Watanabe

Tokai University Research and Information Center, Tokai University, Tokyo, Japan

Takahiro Takamatsu

School of Integrated Design Engineering, Keio University, Yokohama, Japan

Takashi Y. Nakajima

Tokai University Research and Information Center, Tokai University, Tokyo, Japan

(Manuscript received 17 August 2015, in final form 1 February 2016)

ABSTRACT

Variation in surface solar irradiance is investigated using ground-based observation data. The solar irradiance analyzed in this paper is scaled by the solar irradiance at the top of the atmosphere and is thus dimensionless. Three metrics are used to evaluate the variation in solar irradiance: the mean, standard deviation, and sample entropy. Sample entropy is a value representing the complexity of time series data, but it is not often used for investigation of solar irradiance. In analyses of solar irradiance, sample entropy represents the manner of its fluctuation; large sample entropy corresponds to rapid fluctuation and a high ramp rate, and small sample entropy suggests weak or slow fluctuations. The three metrics are used to cluster 47 ground-based observation stations in Japan into groups with similar features of variation in surface solar irradiance. This new approach clarifies regional features of variation in solar irradiance. The results of this study can be applied to renewable-energy engineering.

1. Introduction

Renewable energy systems are looked at as a solution to the problem of global warming because such systems are not based on fossil fuels and emit less carbon dioxide throughout their life cycle. Solar power generation systems, which use solar irradiance as an energy source, are a major type of renewable energy system. Currently, more information about and further understanding of surface solar irradiance features are desired. While the intensity of global solar irradiance at ground level is frequently considered, its variation is also important. Variation on short-term scales is targeted because short-term fluctuation can destabilize electric power systems. Solar irradiance variation at the ground surface affects power plant operation and site selection for solar power plant installations.

Researchers have investigated short-term variations in surface solar irradiance for application to solar power engineering using indices or metrics that characterize the variation or fluctuation in solar irradiance. Lave and Kleissl (2010) and Lave et al. (2012) analyzed the ramp rate (RR) to investigate geographic smoothing effects. Tomson and Tamm (2006) investigated the stability of surface solar irradiance, using absolute values of its increments. Woyte et al. (2007) applied wavelet spectrum analysis to classify fluctuations of solar irradiance. Some researchers have targeted engineering subjects more directly by investigating variations in the output of photovoltaic power generation (e.g., Murata et al. 2009; Marcos et al. 2012). Duchon and O'Malley (1999) suggested that it is possible to determine cloud type during daytime by using solar irradiance data with pyranometer observations at a time resolution of 1 min. They used two statistics to classify cloud type: the mean of a 21-min window and

Denotes Open Access content.

Corresponding author address: T. Watanabe, Tokai University Research and Information Center, Tokai University, 4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan.
E-mail: nabetake@ees.hokudai.ac.jp

DOI: 10.1175/JAMC-D-15-0227.1

the corresponding standard deviation. Their results suggest that multiple variables are needed to characterize the short-term variation of solar irradiance.

Regional features of variation in solar irradiance are also interesting, and such information is important. One major method of clarifying regional features in meteorology is cluster analysis, which is important not only for solar irradiance research but also for other meteorological data (Wilks 2011). Cluster analysis has been applied to solar irradiance data in some previous studies. Yoshida and Kikuchihara (1989) classified regions in Japan according to their solar irradiance magnitude. They used a large amount of monthly surface solar irradiance data, dividing observation sites into five major regions and further dividing those regions into subclusters. Diabate et al. (2004) also considered the magnitude of the surface solar irradiance over all of Africa. Zagouras et al. (2013) proposed a clustering method for high-spatial-resolution data from geostationary satellites. These previous works employing cluster analysis are based on solar irradiance magnitude. Zagouras et al. (2014a,b) also used cluster analysis to consider the variability of solar irradiance. Their unique method provides useful information about regional features of variability in surface solar irradiance.

We aim to clarify regional features of variation of solar irradiance at the ground surface by addressing two main items in this study: the evaluation of variability in surface solar irradiance and the classification of ground-based observation sites in Japan considering their variation features of solar irradiance. These issues are meaningful not only for renewable energy engineering but also for meteorology and climatology. We use solar irradiance data for Japan only; however, our approach can be applied to similar research anywhere.

The remainder of this paper is organized as follows. Section 2 describes the data used. Section 3 describes the method used in this study. Section 4 introduces sample entropy as a new metric for evaluating the variability of solar irradiance. Section 5 discusses the physical interpretation of sample entropy as it relates to variation in solar irradiance. Section 6 presents the results of cluster analysis, and section 7 summarizes this study.

2. Data

In this research, we have used surface global solar irradiance data maintained by the Japan Meteorological Agency (JMA; JMA 1996). Solar irradiance is defined as the accumulated value for 1 min of data sampled at 10-s intervals (Ohtake et al. 2015), and the data time resolution is 1 min. The data period is 5 yr, from 2010 to 2014. We selected 47 observation sites for which 5 yr of consecutive data were available (Fig. 1). Pyranometers at most stations were replaced in the middle of 2011 (Ohtake et al. 2015). JMA performs quality control and routine maintenance of the equipment.

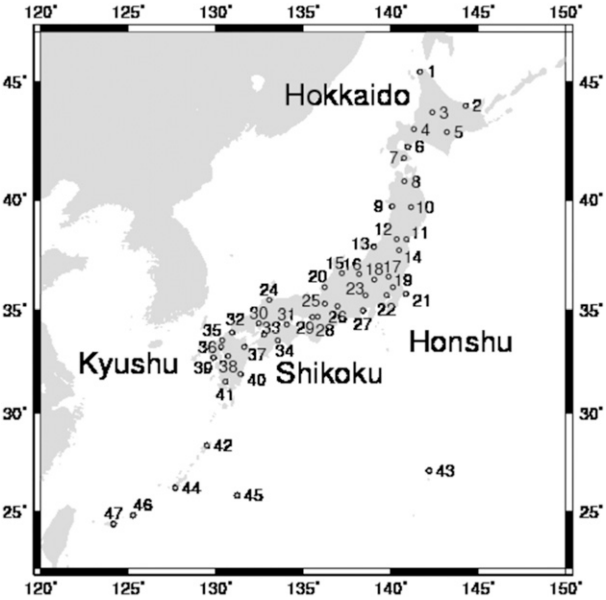

FIG. 1. Locations of observation stations.

We used the clearness index (CI) as defined by Woyte et al. (2007). The CI at time t is defined as the ratio of the observed solar irradiance at the surface I_g to downward irradiance at the top of the atmosphere I_t, as follows:

$$\mathrm{CI}(t) = I_g(t)/I_t \cdot (t).$$

Here, downward irradiance at the top of the atmosphere is calculated as

$$I_t(t) = I_0 E(t) \cos Z(t, l),$$

where I_0 is the solar constant (1353 W m^{-2}), $E(t)$ is the eccentricity factor at time t, and $Z(t, l)$ is the solar zenith angle at time t and latitude l.

Application of the CI significantly reduces the effect of the diurnal and annual cycles on solar irradiance data (Fig. 2). The impact of differences in measurement site latitude on the magnitude of solar irradiance is also reduced. If only cloud effects are considered, large CI values correspond to less reduction of solar irradiance at ground level on cloud-free days. Days with small CI values are overcast. Changes in mean CI are not always linearly related with solar irradiance reductions because of clouds. The direct component of solar irradiance is reduced when cloud thickness increases, but clouds, especially thin clouds, strengthen the scatter component. The CI is thus an indicator of the availability of surface solar irradiance at the observation site. The mean of monthly CI for all

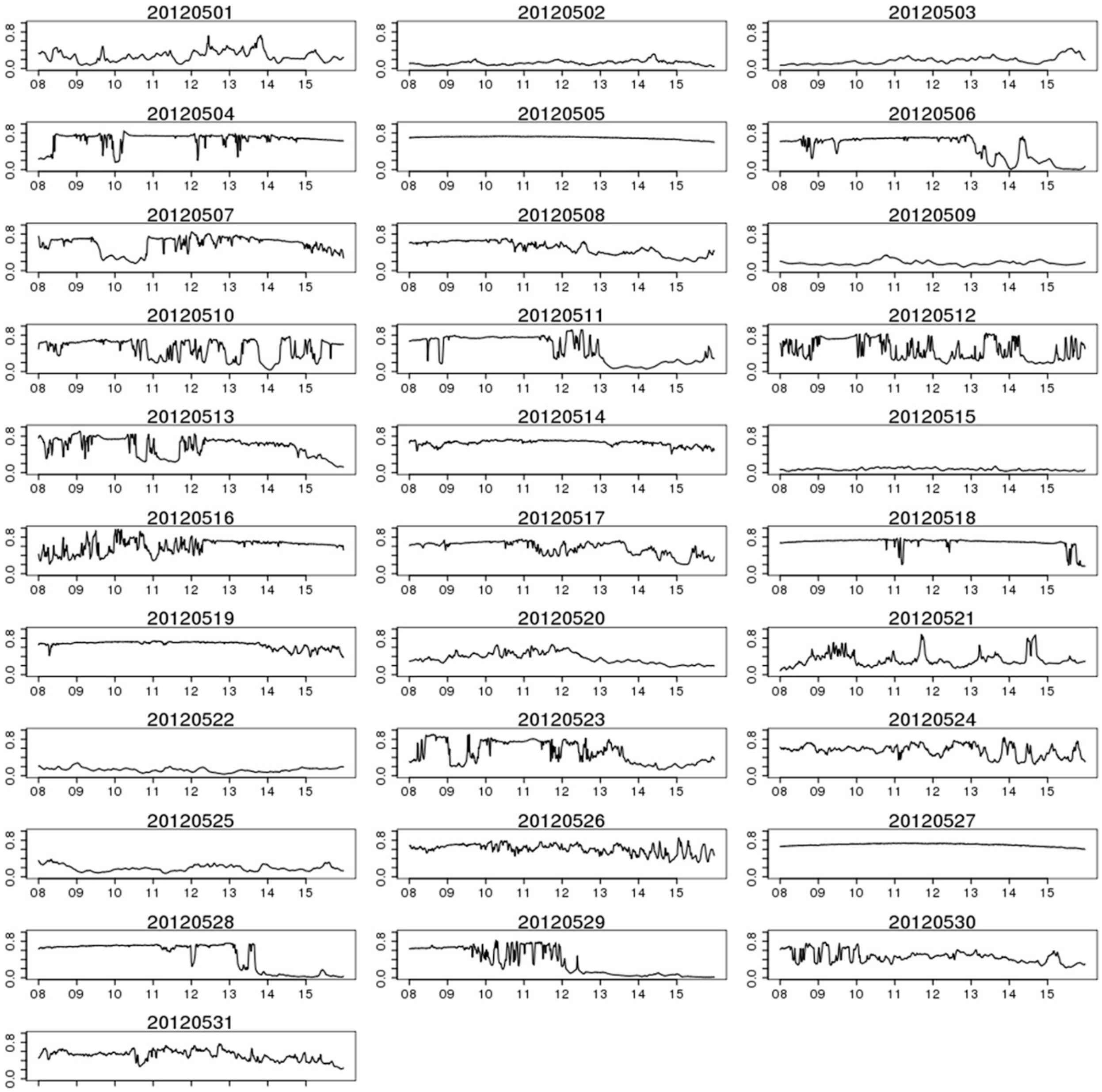

FIG. 2. Time series of CI at the Shizuoka site during May 2012. Shizuoka is station 27 in Fig. 1. The horizontal axis represents the hour of the day (JST).

sites over the 5-yr study period is about 0.443, and its maximum among all sites during this period is 0.6681.

We also use one other weather condition value: the daily mean cloud cover. Cloud cover is a visual observation used to investigate sky conditions related to variations in surface solar irradiance, and is defined as the ratio of cloud coverage over the whole sky. The range of cloud cover is from 0 to 10 (in tenths), with smaller (larger) values corresponding to less (more) cloud coverage. These data are available from the JMA website (http://www.jma.go.jp/jma/index.html). The daily cloud cover data at Shizuoka observation station (station 27) are used, with observations performed three times per day at 0900, 1500, and 2100 Japan standard time (JST). The daily value is the mean of the three observations.

3. Methods

a. *Sample entropy*

The mean and standard deviation are often used to characterize surface solar irradiance. However, these

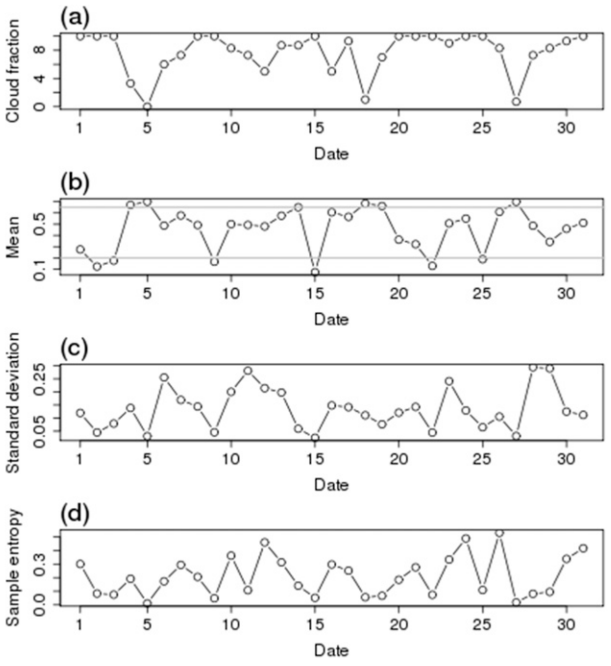

FIG. 3. Daily values of (a) cloud cover, (b) mean, (c) standard deviation, and (d) sample entropy for the CI in May 2012 at the Shizuoka station.

two metrics alone are insufficient for characterizing the variation in solar irradiance, because the standard deviation shows the width of a distribution but does not evaluate features of the arrival order of data. For example, the two time series for 6 and 12 May in Fig. 2 have similar values for the mean and standard deviation of the CI (Fig. 3), but their variations are clearly different. Solar irradiance on 6 May varies slowly and abruptly decreases at around 1300 JST, while rapid fluctuation on 12 May continues over the entire day.

We introduce sample entropy as a metric to represent the features of variation in solar irradiance. Broadly, entropy is a measure that evaluates the complexity or regularity of a system. Sample entropy is a complexity measure for time series, developed by Richman and Moorman (2000). Approximate entropy (ApEn), based on the same information theoretic foundation and developed prior to sample entropy (Pincus 1991; Pincus and Goldberger 1994), provides another measure of time series complexity. The key difference is that sample entropy does not consider self-matches, and as a result does not suffer the same bias as ApEn (Richman and Moorman 2000).

Richman and Moorman (2000) investigated theoretical behaviors of sample entropy, presented as a mathematical definition in appendix A. In brief, sample entropy is the negative natural logarithm of the conditional probability that, given N points, two similar sequences of m points will remain similar to within a tolerance r at the next point. A low sample entropy value indicates a high degree of regularity or low degree of complexity. These measures can be useful tools for analyzing short (but with greater than 100 points) and ''noisy'' time series data (Pincus and Goldberger 1994; Richman and Moorman 2000).

Sample entropy is mainly developed and applied to physiology research, and there have been only a few studies applying sample entropy to meteorological and climatological research. Li et al. (2006) applied sample entropy to daily temperature data at stations over southwest China to investigate climatic divisions in the transition zone between Indian and Chinese monsoons. They interpreted regions with lower sample entropy as being influenced by a homogenous climatic environment, namely, the Indian monsoons. Regions with higher sample entropy were interpreted as complex environments where the two regional monsoons interact.

Theoretical discussion of sample entropy is not the major focus of this study. In the next two sections, we instead show what ''complexity'' denotes in analyzing solar irradiance data and the relationship between sample entropy and physical qualities relating to the variability of solar irradiance.

b. *Wavelet spectrum analysis*

Wavelet analysis is a useful method for investigating localized power variations in time series. The wavelet spectrum analysis in this study is based on Meyers et al. (1993) and Torrence and Compo (1998), and the mathematical basis for analysis is described in appendix B. Solar irradiance varies over the course of a day, and its variation features change several times a day, which means time series of solar irradiance and the CI are nonstationary. We therefore perform wavelet analysis to investigate power spectrum distributions of the CI. Wavelet spectrum analysis of the CI time series is presented in section 5.

c. *Cluster analysis*

Cluster analysis is a useful method for dividing large datasets into groups (Wilks 2011). In this study, we apply Ward's minimum variance method (hereinafter Ward's method). The theoretical basis and methodology are mainly according to Wilks (2011) and Murtagh and Legendre (2014). Ward's method is a hierarchical clustering approach. The similarity measure in Ward's method is based on the sum-of-squares distance within a cluster, called the within-cluster variance. The clustering process begins with each station belonging to its own cluster. At each clustering step, called an ''analysis stage,'' the sum of within-cluster variances over all variables is minimized as clusters are merged. The

Mahalanobis distance is applied to measure the distance between clusters, because multivariable data are classified and correlation effects between variables need to be reduced. We used the R programming language with version 3.1.1 of the ''stat'' package for cluster analysis processing.

To determine how many clusters the set of observation sites should be divided into, we consider changes in within-cluster variance summed over clusters as a function of the analysis stage. Larger changes in the within-cluster variance before and after a clustering step indicate significant separation between two merged clusters. When such a significant change is seen at later clustering stages, the clustering procedure should be terminated, as the number of clusters just before this significant change is a suitable number of clusters. The Caliński–Harabasz ''pseudo-F'' statistic is also used (Caliński and Harabasz 1974; Fovell and Fovell 1993). This statistic is given by the formula

$$\text{Pseudo-F} = (A/W)[(n-k)/(k-1)],$$

where A and W are the among- and within-cluster variances, respectively; n is the number of objects; and k is the number of existing clusters. As suggested by Caliński and Harabasz (1974), cluster numbers should be chosen according to whether the pseudo-F has an absolute or local maximum. Cluster analysis is discussed in section 6.

4. Analysis of solar irradiance metrics

In this section, we address what the ''complexity'' of the solar irradiance data indicates. Sample entropy is a function of three parameters: N, r, and m. Following Pincus (1991) and Richman and Moorman (2000), m is set to 2 and r to 0.25 times the standard deviation. The mean of the standard deviations over all sites for 5 yr is used, the value of which is 0.156, resulting in an r value of 0.039. Solar irradiance data for the 480 min between 0800 and 1559 JST are used, so N is set to 480.

This paragraph discusses the evaluation of variation in the CI using the three metrics—mean, standard deviation, and sample entropy—for cloud-free, overcast with thick clouds, and partially cloudy conditions. We analyzed solar irradiance data from different sky conditions at the Shizuoka station (station 27 in Fig. 1) during May 2012. Figure 2 shows daily time series of the CI, and Fig. 3 shows daily values of the cloud cover, mean, standard deviation, and sample entropy for the CI. The daily mean cloud cover was a small 0.7 tenths on 27 May, a cloud-free day. The mean CI was 0.70, and the standard deviation a very small 0.03. The sample entropy value (0.02) was small under these clear-sky conditions, which indicates that irradiance during a cloud-free day varies regularly. There are two major types of solar irradiance variation on cloudy days: variation where high frequencies are dominant and that where low frequencies prevail. The time series for 22 May is an example of low-frequency variation. Such variation occurs when thick clouds cover the sky throughout the day; then the cloud cover becomes large (10.0 tenths). Small mean CI (0.129) and standard deviation (0.045) are seen on such a day. An example of the other variation type is 26 May, which was also cloudy throughout the day with a large cloud cover of 8.3 tenths. The CI mean and standard deviation are 0.609 and 0.105, respectively. The sample entropy is smaller for the day with slower variability (0.07) than for the day with fast fluctuation (0.53).

Sample entropy is consistently low for days with low standard deviation and high or low mean CI, which indicates variation in cloud-free days (5 and 27 May) and cloudy days with thick clouds (2, 3, 9, 15, 22, and 25 May) can be recognized only from the mean or standard deviation of the CI. For days with moderate CI (roughly speaking, mean values between 0.2 and 0.65; see gray lines in Fig. 3b), the relationship between standard deviation and sample entropy exhibits a complicated pattern of behavior. For example, the time series for 6 May and 12 May in Fig. 2 are for moderate CI days and show almost the same values for standard deviation. However, differences in the variation features are apparent; fluctuation over 12 May is more rapid than that over 6 May. Values of sample entropy in these two cases are 0.1733 and 0.4604, respectively, indicating that rapidly fluctuating time series have higher complexity. Another pair of days with nearly the same mean and standard deviation values is identified on 17 and 24 May, and these two show different values for sample entropy, namely, 0.2517 and 0.4892. The three largest values for sample entropy are seen on 26, 24, and 12 May, in order of decreasing sample entropy (0.53, 0.49, and 0.46, respectively) with respective standard deviations of 0.105, 0.129, and 0.214. It may be difficult to recognize the differences in variation between the pair only from the two metrics of the mean and standard deviation, but using sample entropy clarifies the differences in variation. This is because the arrival order of the data is crucial for sample entropy, but is not considered in calculations of the standard deviation.

We clarify the relation between the three metrics by using a simultaneous scatter diagram (Fig. 4). The relation between mean and standard deviation acts like a second-order function, but its peak is located on the right side. When the CI is larger, the standard deviation is smaller, which tends to occur under cloud-free

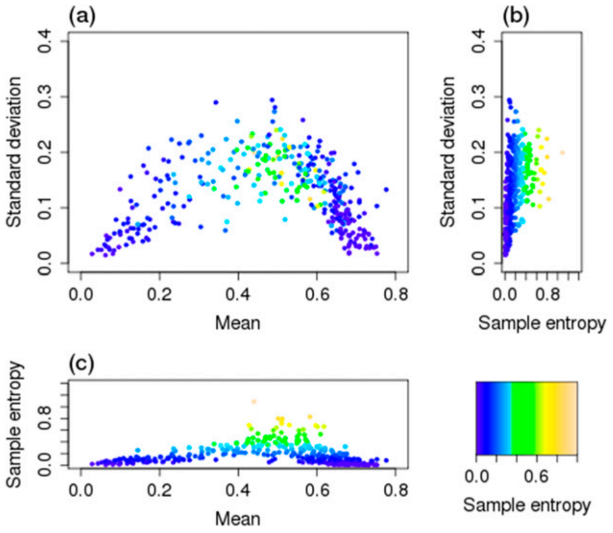

FIG. 4. Scatterplot of three variables at the Shizuoka station in 2012; (a) mean CI values are on the horizontal axis, standard deviations on the vertical axis, and sample entropy is represented by color. The color bar on the right side of the figure bottom is for sample entropy. (b) Scatterplot of sample entropy (horizontal axis) and standard deviation (vertical axis). (c) Scatterplot of mean (horizontal axis) and sample entropy (vertical axis).

conditions. The standard deviation is largest when CI has a moderate value. A smaller CI accompanies smaller standard deviations, which are expected on overcast days with thick clouds. The relation between the mean and standard deviation is similar to that in Duchon and O'Malley (1999), though the data range for processing is different; they used a 21-min window width. On both sides of the second-order-function-like distribution, sample entropy always has a small value when corresponding to regularity in the variation of solar irradiance. Larger sample entropy values are seen only around the peaks of the secondary second-order-function-like distribution centered at around a 0.5 mean and a 0.2 standard deviation, corresponding to the moderate CI zone (Figs. 4a–c). However, the sample entropy is not always large when the mean CI is moderate and standard deviation is larger; it is distributed across a broad range from small to large values. Such a distribution of sample entropy in the moderate CI zone reflects the difference in the manner of the variation.

These relations between the three variables are seen at all sites, but several differences in shape and distribution are seen between stations (Figs. 4–6). For example, more points are plotted on the right side of the second-order-function-like distribution at Shizuoka station, while the distribution at Naha station (station 44; Fig. 5) has more points on the left side, and points are widely distributed over the moderate mean zone at Sapporo station (station 4; Fig. 6). The sample entropy peak is much larger at the Naha site than at the other two sites. These differences indicate that there are regional features of variation in solar irradiance.

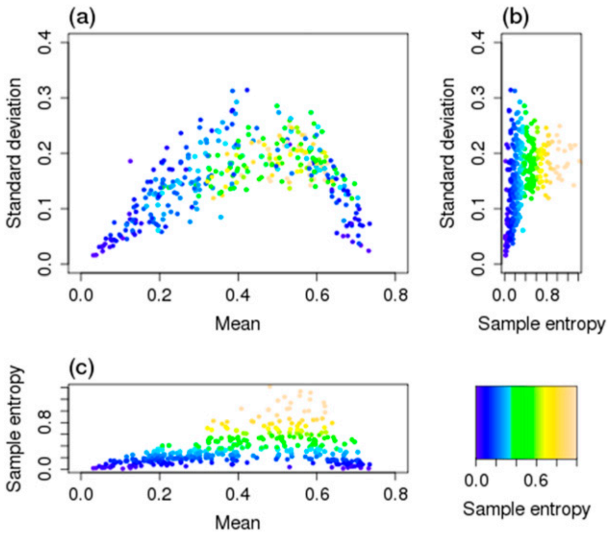

FIG. 5. As in Fig. 4, but for Naha station (station 44).

Table 1 shows a Pearson's correlation coefficient matrix between the three metrics, calculated using monthly data from all 47 sites over 5 yr. The mean shows less correlation with the other two variables, and correlation between the standard deviation and sample entropy is large.

5. Physical interpretation of sample entropy related with variation in solar irradiance

To obtain the physical basis for sample entropy within the context of the solar irradiance, two

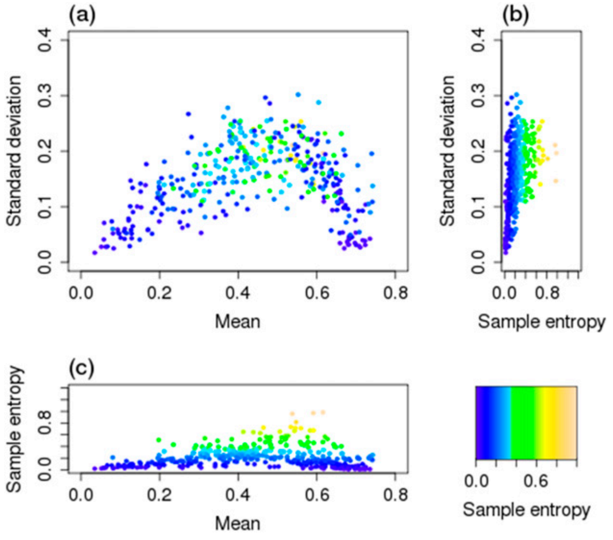

FIG. 6. As in Fig. 4, but for Sapporo station (station 4).

TABLE 1. Correlation matrix of three variables. Each variable is standardized before processing, such that the mean is 0 and standard deviation is 1.

	Mean	Std dev	Sample entropy
Mean	1.000		
Std dev	−0.136	1.000	
Sample entropy	−0.013	0.282	1.000

variability metrics are investigated and compared with sample entropy.

One is wavelet spectrum distribution, which provides information about the variation of the time series as a function of a frequency. The other is the ramp-rate distribution, which represents the change in intensity with respect to time.

a. *Wavelet spectrum distribution*

Pincus and Goldberger (1994) suggested that power spectrum distributions are related to sample entropy values. We performed wavelet spectrum analysis and examined daily mean wavelet power spectra for May 2012 at Shizuoka station (Fig. 7). The significance test is based on the monthly red noise model of a lag-1 autoregressive [AR(1)] process (Torrence and Compo 1998) and the monthly mean variance of the CI time series.

When comparing the power spectrum distribution for time series with almost the same mean CI and standard deviation (6 and 12 May), it is seen that the power spectra are significant for periods shorter than about 120 min. The magnitude of spectra shorter than 50 min on 12 May is stronger. These two time series have almost the same value for the standard deviation, so the total power spectra are nearly the same. The time series on 12 May, which has larger sample entropy, has more spectrum power on the shorter-period side. The other pair, 17 May and 24 May, also shows different spectrum distributions; the time series on 24 May with larger sample entropy has more power during shorter periods from 50 to 100 min.

We suppose that sample entropy is related to the power spectrum distribution and that shorter-period fluctuations in particular affect the sample entropy strength. To confirm this, we used scale-averaged wavelet spectra for periods shorter than 120 min as a metric to represent the shorter-period variation strength. The sample entropy correlates with the scale-averaged wavelet spectra for shorter periods more than does the total wavelet power spectrum over a day, which is equal to the variance of the time series (see appendix B) (Fig. 8a).

The correlation coefficients between the sample entropy and scale-averaged wavelet power spectra for 2012 are 0.756 and 0.654 by the Spearman and Pearson methods, respectively (Fig. 8b and Table 2). High positive correlations indicate that sample entropy is related with spectrum distributions at shorter periods.

As Fig. 8c shows, we also use the ratio of scale-averaged wavelet spectra to the total wavelet power spectrum in a day. Sample entropy tends to be large when scale-averaged wavelet power spectra for shorter periods are large. The two correlation coefficients between sample entropy and the ratio of scale-averaged wavelet spectra are 0.548 (Spearman) and 0.471 (Pearson) (Fig. 8d).

When the standard deviation is large and shorter-period variation is small, for example, on days 6, 28, and 29, the sample entropy is smaller. Solar irradiance on these days can increase or decrease abruptly and the decline is step wise on one occasion (Fig. 2). Sample entropy is insensitive to such step-wise changes in the CI. Such variation corresponds to the bottom of the moderate CI zone in Fig. 4a. When the standard deviation is moderate but the ratio of shorter-period variations is large, such as on days 26, 30, and 31, sample entropy is larger. However, when the standard deviation is too small, the sample entropy becomes small, even in cases where the shorter-period variation is large, such as on days 15 and 18. These time series correspond to both sides of the second-order-function-like distribution (Fig. 4a). This is likely due to the parameter r of the sample entropy, because r works as a noise cutoff filter (Pincus and Goldberger 1994). Within the context of variation in solar irradiance, we should use not "noise" but rather ramp size, because it is caused by radiance fluctuations from clouds and sky integrated over the sensor response angles. This is explained in the next subsection.

b. *Ramp-rate distribution*

The RR represents the change in the magnitude with respect to time. The ramp rate at time t is defined as the difference between sequences of the CI, that is, CI at times t and $t + 1$:

$$\mathrm{RR}(t) = [\mathrm{CI}(t + dt) - \mathrm{CI}(t)]/dt,$$

where dt is the temporal interval of the data.

According to Lave et al. (2012), the RR distribution is plotted showing the cumulative frequency of occurrence of the absolute value of the RR (Fig. 9). The distribution shape is related to the manner of variation in the CI for the day. The cumulative frequency increases slowly when the CI fluctuates rapidly and strongly (e.g.,

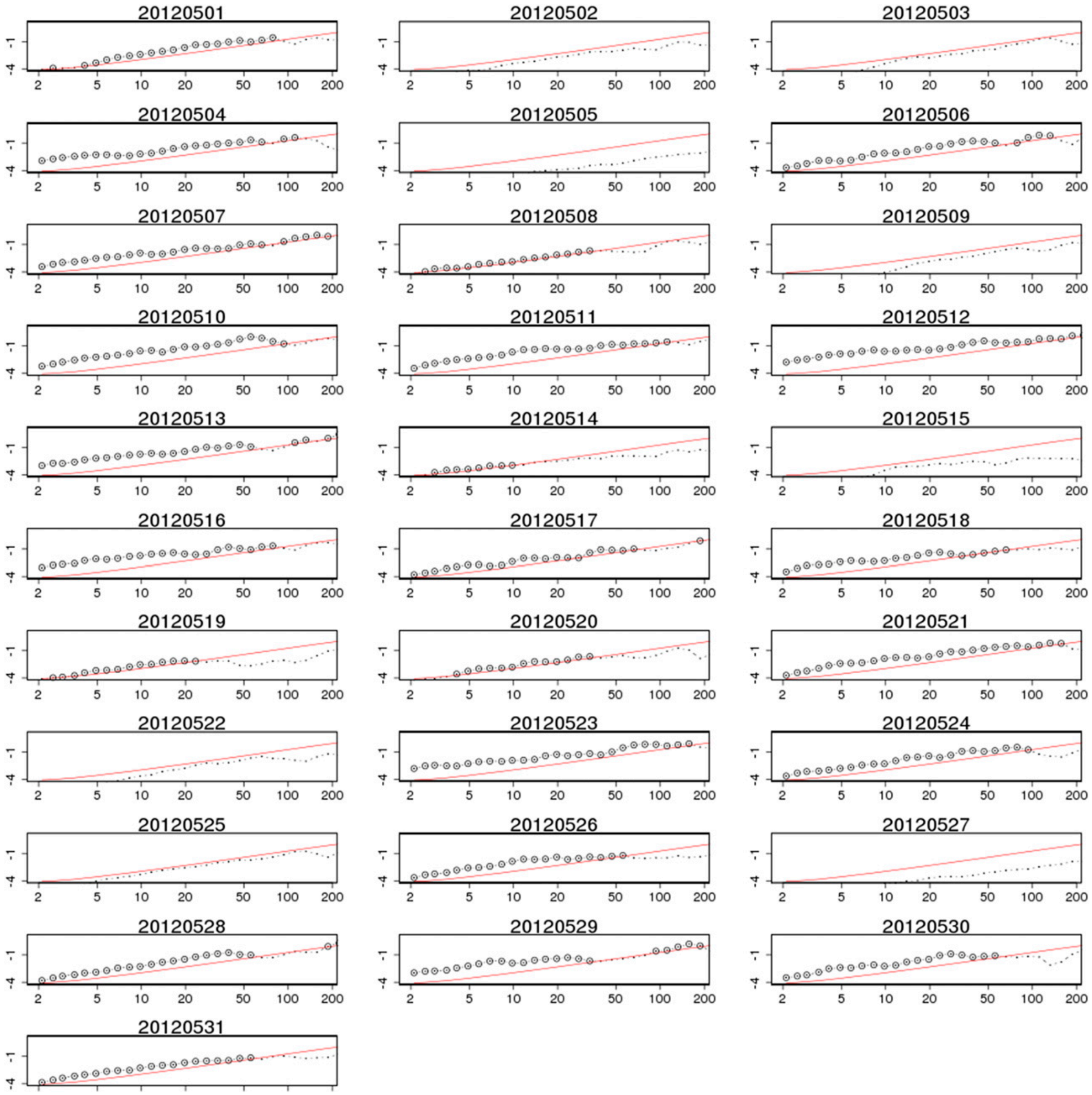

FIG. 7. Temporally averaged power spectra as a function of period over a day. The horizontal axis represents the period (min). The vertical axis represents the logarithm to the base 10 of the power spectrum. The red line represents 95% confidence levels. Circles indicate that the power spectrum at a frequency is significant.

12 May). However, the cumulative frequency tends to become saturated at low RR on low-variability days (e.g., 22 May and 27 May). To condense the RR distribution down to a single value, the RR at a given percentile, or quantile, is used (Lave et al. 2012). The locations of the RR at the 70th, 80th, and 90th percentiles (hereinafter, RR70, RR80, and RR90, respectively) also change, reflecting the shape of the cumulative RR distribution (Fig. 9).

Two time series of sample entropy and the RR90 for May 2012 synchronize well (Fig. 10a). The correlation between sample entropy and the RR90 over 2012 is 0.736 and 0.856 by use of the Spearman and Pearson methods, respectively (Fig. 10b and Table 2). A pair of days with nearly the same mean and standard deviation values—17 May and 24 May, which are clearly separated using sample entropy—shows a difference in the RR distribution; each percentile of the RR on 24 May shifts

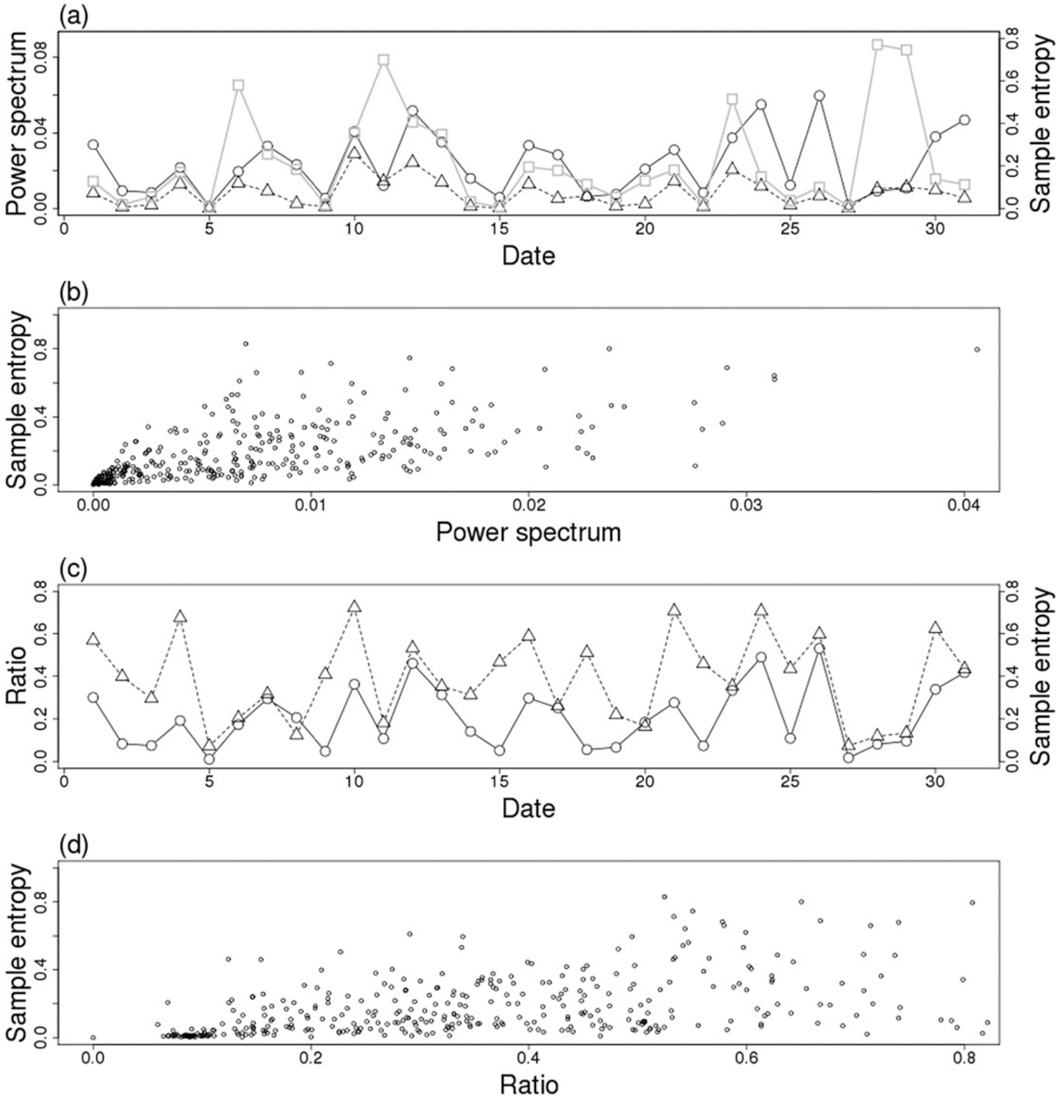

FIG. 8. (a) Scale-averaged wavelet spectra for periods shorter than 120 min for May 2012 at the Shizuoka station (dotted line with triangles). The black line with circles corresponds to the sample entropy and the gray line with squares to the total wavelet power spectrum (square of standard deviation). (b) Scatterplot of scale-averaged wavelet spectra for periods shorter than 120 min (horizontal axis) and sample entropy (vertical axis) for the year 2012 at the Shizuoka station. (c) Ratio of scale-averaged wavelet spectra for periods shorter than 120 min to the total wavelet power spectrum (dotted line with triangles). Sample entropy is plotted as a black line with circles. (d) As in (b), but for the ratio of scale-averaged wavelet spectra for periods shorter than 120 min (horizontal axis) and sample entropy (vertical axis).

to a larger RR value than on 17 May. This indicates that variation on 24 May has stronger variability.

When the parameter *r* of the sample entropy is changed, the correlation between the sample entropy and RR also changes (Table 3). When *r* becomes larger, the quantile value of the RR, which is most strongly correlated with sample entropy, shifts to a larger value. This correlation analysis indicates that *r* of the sample entropy serves as the threshold for the solar irradiance ramp.

c. Conclusions from physical interpretation of sample entropy

From analysis and discussion of the relation between sample entropy and the wavelet spectrum and RR distribution metrics, this section can be summarized by saying that sample entropy is a metric representing the manner of fluctuation in solar irradiance. The results of the correlation analysis indicate that sample entropy, the wavelet spectrum, and the RR distribution have

TABLE 2. Pairwise correlations between among sample entropy, wavelet spectrum, and RR90. Left of the solidus (/) is the Pearson correlation coefficient; right is the Spearman correlation coefficient.

	Sample entropy	Wavelet	RR90
Sample entropy	1.000		
Wavelet	0.654/0.756	1.000	
RR90	0.736/0.856	0.880/0.903	1.000

similar roles in evaluating the variation in solar irradiance. The advantage of using sample entropy is that it is a single value. The RR distribution and wavelet spectrum are, respectively, functions of the RR and frequency, so some procedure is necessary to obtain a single value from these functions.

However, some questions about sample entropy remain. One is how to select parameters *r* and *m* of the sample entropy. In this study, these parameters are determined following previous works (Pincus 1991; Richman and Moorman 2000). For example, Lake et al. (2002) proposed how to mathematically select *r* and *m*. More detailed discussion of sample entropy will be helpful for applying the sample entropy to investigations into solar irradiance.

6. Results and discussion of cluster analysis

a. *Geographic assessment of solar irradiance clustering*

We performed cluster analysis using three metrics: the mean, standard deviation, and sample entropy, which, respectively, represent the availability of solar irradiance at the surface, the strength of the variation, and the manner of fluctuation. Before performing cluster analysis, each metric is standardized such that its mean value over the year becomes 0 and the corresponding standard deviation becomes 1. Standardization is performed for data from every year, which reduces the interannual variation and the influence of the equipment replacement (Ohtake et al. 2015). The effect of the weak correlation between standard deviation and sample entropy (Table 1) can be reduced using the Mahalanobis distance.

Note that there are several limitations in the cluster analysis. Because the global solar irradiance data cover only a 5-yr period, obtaining robust climatological features from the results is difficult. For example, the annual mean El Niño–Southern Oscillation (ENSO) indices [the difference in area-averaged sea level pressure between Tahiti and Darwin, available from the JMA website (http://www.data.jma.go.jp/gmd/cpd/data/elnino/index/soi.html)] from 2010 to 2014 are 1.13, 1.35, 0.10, 0.57, and −0.13, which indicate that La Niña is dominant over these five years. Then, the influence of one phase of the ENSO on weather over Japan may be emphasized for this data period. The locations and number of stations may also impose limitations; observation sites are coarsely distributed over mainly urban areas, and variation features may be related to the urban environment. Also, it is difficult to determine appropriate geographic boundary lines for dividing regions into clusters, because of the coarse location of the observation stations. Despite these limitations, we emphasize that cluster analysis using the three metrics of variation in the CI is a new approach to clarifying regional features of variation in surface solar irradiance.

Figure 11 shows change in the within-cluster variance during the analysis stage, and Fig. 12 shows the result of cluster analysis for the three- and six-cluster cases. The height along the vertical axis in Fig. 11 indicates the change in within-cluster variance before and after merging two clusters in the analysis stage. Because of the criterion of within-cluster variance change, the number of clusters is determined to be 3, 6, or 9. However, the Caliński–Harabasz pseudo-F statistic has local maxima in the three- and six-cluster cases, so the discussion below mainly focuses on the cases of three and six clusters. Clusters are designated by A, B, and C in the three-cluster case, and by numbers 1–6 in the six-cluster case.

In the three-cluster case, sites on large islands are divided into two major clusters: A and B (Fig. 12a). Cluster A comprises sites in northern Japan and islands in the Sea of Japan. The other cluster of large islands, cluster B, comprises sites on Kyushu, Shikoku, and Pacific-side Honshu. Cluster C, which is composed of sites on small islands, is clearly separated from sites on large islands. This is supported by the fact that clusters of small islands merge with those of larger islands during the final clustering stage.

In the six-cluster case, each major cluster in the three-cluster case is separated into two clusters (Fig. 12b). Cluster A is divided into clusters 1 and 2. Cluster 1 includes sites on Hokkaido, except for the northernmost site, and northern Honshu on the Pacific side. Cluster 2 comprises sites on Honshu and Hokkaido along the Sea of Japan. Cluster B is divided into eastern Honshu (cluster 3) and other sites in Kyushu and western Honshu (cluster 4). Sites on the small islands of cluster C are further separated into two clusters: 5 and 6. Note that cluster 5 involves only two sites.

Most clusters in each case are composed of sites with close geographical proximity. The northernmost observation site at Hokkaido (station 1 in Fig. 1) may appear to be apart from the other sites of cluster 2. All sites of cluster 2 are located on the Japan Sea side, though

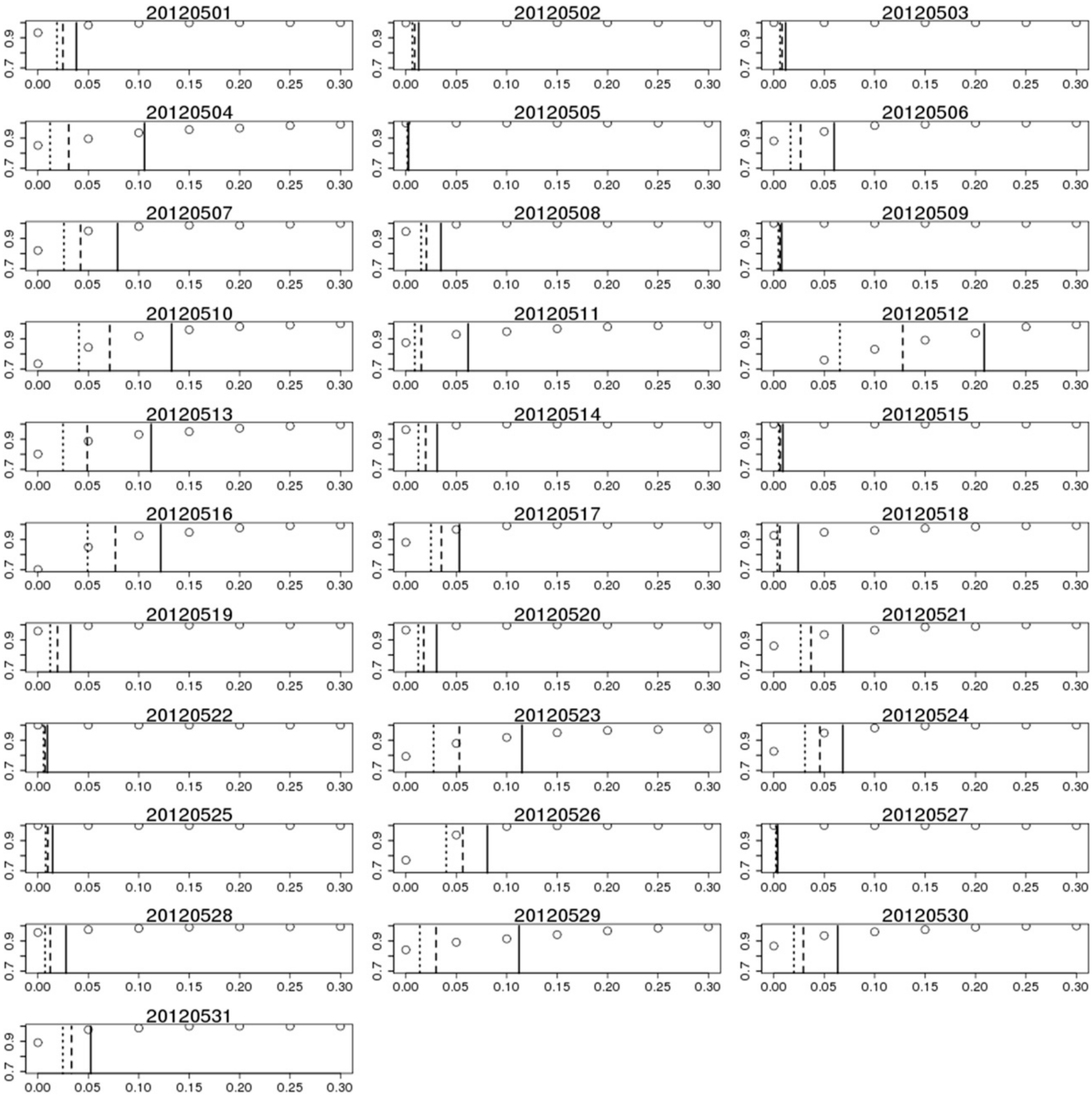

FIG. 9. The cumulative RR distribution for each day in May 2012 at the Shizuoka station. The cumulative distribution along the vertical axes is the ratio of occurrences of the RR value smaller than or equal to the RR value on the horizontal axes against the total number in a day. The dotted, broken, and solid lines are, respectively, drawn at the 70th, 80th, and 90th percentiles of the cumulative distribution.

station 8 is a bit inland. So cluster 2 seems to characterize the variation in solar irradiance on the Japan Sea side. The separation of station 1 is likely due to the coarse distribution of the locations of the observation sites.

b. *Solar irradiance metrics as a function of cluster*

The results of cluster analysis using the three metrics clarify characteristics of the monthly mean value for three metrics of CI type in solar irradiance (Fig. 13). The benefit of using multiple metrics to examine variability is that we can identify regional features of variation from several aspects, as part of our discussion of the six-cluster case below.

One of the two major clusters of large islands sites, cluster A, includes clusters 1 and 2. These two clusters have similar characteristics for standard deviation and sample entropy, in that both variables are minimized in late spring (Figs. 13a,b). However, their means are

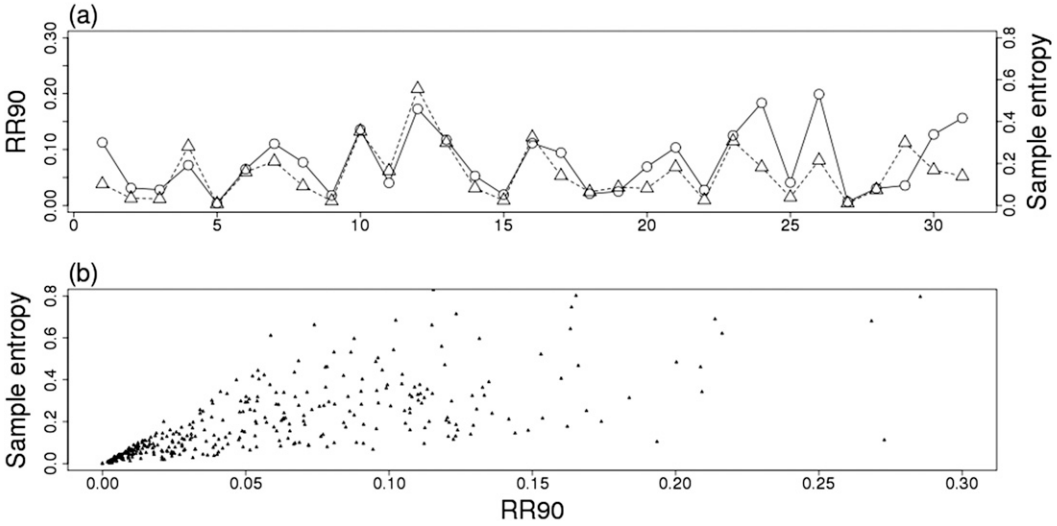

FIG. 10. (a) Ramp rate at the 90th percentile of the cumulative distribution (RR90) for each day in May 2012 at the Shizuoka station (dotted line with triangles). The solid line with circles corresponds to the sample entropy and the dotted line with triangles to the RR. (b) Scatterplot of the RR90 (horizontal axis) and sample entropy (vertical axis) for 2012 at the Shizuoka station.

clearly different. The mean CI of cluster 2 is smaller in winter and increases in summer (Fig. 13b). It thus seems that late spring and early summer are suitable times to use solar irradiance energy at sites in cluster 2. In contrast, cluster 1 has its largest mean CI in February, gradually decreasing to a minimum in December. The standard deviation and sample entropy are larger during the winter (Fig. 13a).

Observation sites in cluster 3 have relatively weaker and more regular variation than do other clusters. The availability of solar irradiance is high in winter, but the strength and complexity of variation is small. The mean CI of cluster 3 is maximal in January and has two minima: in June and October (Fig. 13c). Standard deviation and sample entropy have one peak around September and June, respectively, but are weak in the other seasons. Surface solar irradiance at cluster 3 sites varies with strong and rapid fluctuations in summer and autumn, but the variation in solar irradiance is calm in winter. While the mean value of cluster 4 has a notable minimum in June, the availability of solar irradiance does not significantly change in the other months (Fig. 13d). Both the standard deviation and sample entropy from March to May are smaller, so solar energy can be stably available in spring. In summer, the July-to-September variability and complexity are larger, despite the mean value also being large.

Variation features at small island sites (clusters 5 and 6; Fig. 12b) are characterized by the high availability of solar irradiance in summer. However, when comparing clusters 5 and 6, differences in variation features are clearly seen (Figs. 13e,f). The mean and standard deviation of cluster 5 have two peaks. However, sample entropy has one significant maximum in October. The availability of solar irradiance is large at sites in cluster 5, especially during the summer, but strong and rapid variation occurs in October. Sample entropy values at sites on small islands are larger than those for large island sites; in particular, sites on the southwestern small islands (cluster 6) show a significant peak in July. While large solar irradiance is available in summer, the variation tends to fluctuate both strongly and rapidly. In winter, solar irradiance availability and complexity are considerably reduced at the stations in cluster 6.

c. Discussion

The causes for regional features of solar irradiance are an interesting question. As suggested by Duchon and O'Malley (1999), variation in solar irradiance on shorter temporal scales is related to cloud type. To seek the answer, a combination of two types of information is necessary: features of shorter temporal-scale variation

TABLE 3. Correlation between sample entropy and percentile of the RR distribution. Left of the solidus (/) is the Pearson correlation coefficient; right is the Spearman correlation coefficient.

		RR70	RR80	RR90
Sample entropy	$r = 0.04$	0.911/0.959	0.846/0.931	0.736/0.856
	$r = 0.08$	0.955/0.962	0.926/0.965	0.842/0.925
	$r = 0.12$	0.954/0.944	0.951/0.964	0.892/0.949

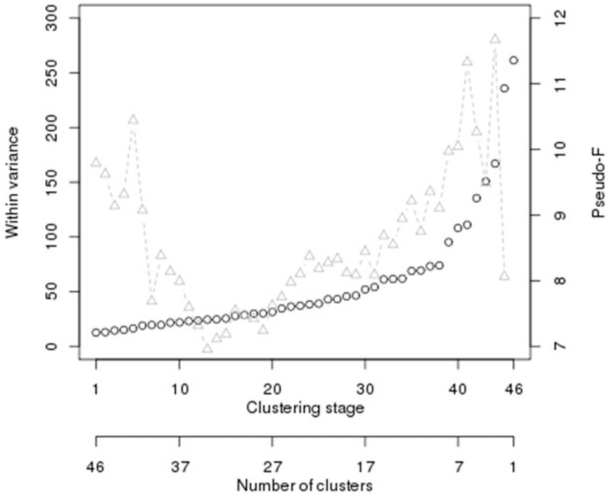

FIG. 11. Change in within-cluster variance during the clustering stage (black solid line with circles; left vertical axis). Top and bottom numbers on the horizontal axes are the clustering stage and number after merging. The right axis represents the Caliński–Harabasz pseudo-F statistic during the cluster stage (gray dashed line with triangles; right vertical axis). The pseudo-F statistic is not defined at the final stage of cluster analysis.

in the solar irradiance and classification of the cloud type. Information about cloud type is obtained from satellite and ground-based observations. In this study, three metrics are computed based on the daytime CI time series. If the same method is applied to time series on intraday scales, we might obtain information about the variation related to cloud type.

Understanding regional variation features in more detail requires addressing the question of how many variables are appropriate. As mentioned in section 1, there are some metrics to represent the variability of solar irradiance. Considering the correlation analysis in section 5, it is possible that several metrics correlate with each other, causing redundancy. We may lack metrics to represent meaningful features of variability. Such questions will be addressed in a future study.

Our goal is to give the information about variation in surface solar irradiance that is desired in solar energy engineering. This study can help inform grid operators of regions that may more frequently experience ramping conditions, and which regions may have smoother and more consistent power delivery. Regional features of variation in solar irradiance are likely to identify needs for devices or increased ancillary power generation in a given region. Additionally, the efficiency and durability of electric devises may be able to be improved by considering regional variation features in solar irradiance.

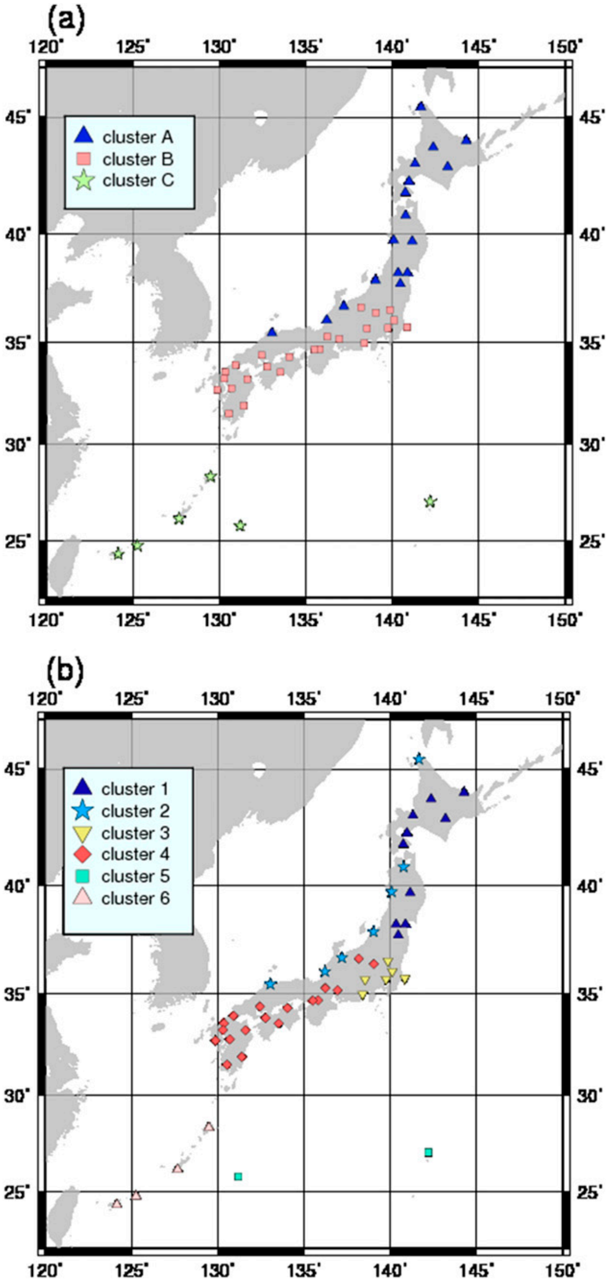

FIG. 12. Results of cluster analysis in the (a) three- and (b) six-cluster cases. Marks represent resultant clusters in each cluster case.

7. Summary

We investigated solar irradiance variation at the surface on daily (whole daytime) temporal scales, addressing two items in particular: evaluation of solar irradiance variation and regional features of variation in solar irradiance. Understanding and offering information related to variations in solar irradiance can be usefully

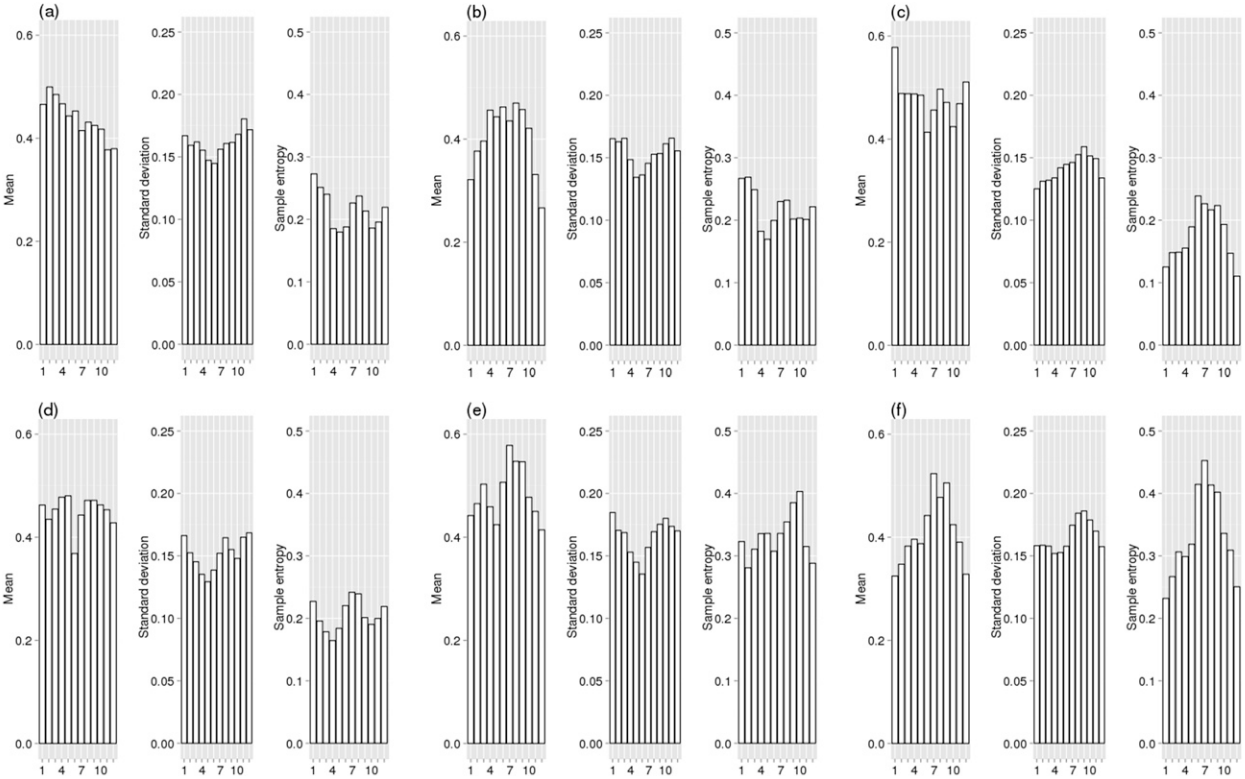

FIG. 13. Monthly means of the three daily variables for each cluster in the six-cluster case: (a)–(f) clusters 1–6, respectively. Each panel includes three graphs showing the mean, standard deviation, and sample entropy of the CI, from left to right. The horizontal axis of each graph represents the months from January to December.

applied to demands from society and to renewable-energy engineering.

We used three variables—mean, standard deviation, and sample entropy—to evaluate the variation in solar irradiance in more detail. These variables, respectively, represent the availability of solar irradiance at the surface, the strength of the variation, and the manner of fluctuation. The observed 1-min surface solar irradiance was divided by the solar irradiance at the top of the atmosphere to reduce the effects of diurnal and annual cycles and latitudinal effects. This radiation variable was analyzed in this investigation.

Because sample entropy has not been well applied to meteorological and climatological research, we addressed what “complexity” denotes in analyzing solar irradiance data and the relationship between sample entropy and physical qualities relating to the variability of solar irradiance. The qualitative features of sample entropy are as follows: sample entropy becomes smaller under clear skies and in overcast conditions with thick clouds. Large sample entropy occurs during daytime with moderate solar irradiance. Large standard deviations also occur with moderate solar irradiance, but do not always indicate large sample entropy. There is a weak positive relation between the standard deviation and sample entropy. Rapid fluctuation of solar irradiance corresponds to larger sample entropy. Sample entropy is insensitive to slow fluctuations in solar irradiance, like abruptly and step-wise variations on one occasion.

We investigated regional features of variation in solar irradiance using these three metrics. Cluster analysis divided 47 sites in Japan into groups with similar features of variation. We found that three- and six-cluster groups are suitable for classifying the 47 sites. From the results of our cluster analysis, the monthly mean of the three metrics can be obtained for each cluster. We can identify regional features of variation from several aspects by employing this approach.

Acknowledgments. This work was supported by the Japan Science and Technology Agency through the CREST/EMS funding program. The computer program for wavelet spectrum analysis used in this study is partly based on code provided by C. Torrence and G. Compo and is available online (http://atoc.colorado.edu/research/wavelets/).

APPENDIX A

Definition of Sample Entropy

Following Richman and Moorman (2000), we define sample entropy as follows. A time series of N points $\{u(j)$: $1 \le j \le N\}$ forms the $N - m + 1$ vectors $x_m(i)$ for $\{i$: $1 \le i \le N - m + 1\}$, where $x_m(i)=\{u(i + k)$: $0 \le k \le m - 1\}$ is a vector of m data points from $u(i)$ to $u(i + m - 1)$. The distance between two such vectors is defined to be $d[x_m(i), x_m(j)] = \max\{|u(i + k) - u(j + k)|$: $0 \le k \le m - 1\}$, the maximum difference of their corresponding scalar components.

We define $B_i^m(r)$ as $1/(N - m - 1)$ times the number of vectors $x_m(j)$ within r of $x_m(i)$ measured with $d[x_m(i), x_m(j)]$, where j ranges from 1 to $N - m$, and $i \neq j$:

$$B^m(r) = (N - m)^{-1} \sum_{i=1}^{N-m} B_i^m(r).$$

Similarly, we define $A_i^m(r)$ as $1/(N - m - 1)$ times the number of vectors $x_m + 1(j)$ within r of $x_m + 1(i)$, where j ranges from 1 to $N - m$, and $i \neq j$:

$$A^m(r) = (N - m)^{-1} \sum_{i=1}^{N-m} A_i^m(r).$$

Here, $B_m(r)$ is the probability that two sequences will match for m points, whereas $A_m(r)$ is the probability that two sequences will match for $m + 1$ points.

We define the sample entropy SampEn as

$$\mathrm{SampEn}(m, r, N) = -\ln[A_m(r)/B_m(r)].$$

Sample entropy has three parameters: m, r, and N, where m is the length of the sequences to be compared, r is the tolerance for accepting matches, and N is the length of the time series.

APPENDIX B

Wavelet Spectrum Analysis

Wavelet analysis in this study is based on the work of Torrence and Compo (1998). Assume a discrete time series $\{x(n)$: $0 \le n \le N - 1\}$ with uniform time spacing δt and a wavelet function $\Psi(\eta)$. The wavelet transform $W_n(s)$ of the discrete time series is defined as

$$W_n(s) = \sum_{n'=0}^{N-1} x(n')\Psi^* \left[\frac{(n' - n)\delta t}{s}\right],$$

where s is the wavelet scale and the asterisk indicates complex conjugation.

A time-averaged wavelet spectrum over the period from n_1 to n_2 can be obtained by

$$\overline{W_n^2}(s) = \frac{1}{(n_2 - n_1 + 1)} \sum_{n_1}^{n_2} |W_n(s)|^2.$$

The scale-averaged wavelet spectrum over scales j_1 to j_2 is defined as

$$\overline{W_n^2} = \frac{\delta j \delta t}{C_\delta} \sum_{j=j_1}^{j_2} \frac{|W_n(s_j)|^2}{s_j},$$

where δ_j and C_δ are constant values related to the wavelet function.

The variance σ^2 of time series x can be related to a wavelet power spectrum by using an equivalent of Parseval's theorem for wavelet analysis:

$$\sigma^2 = \frac{\delta j \delta t}{C_\delta N} \sum_{n=0}^{N-1} \sum_{j=0}^{J} \frac{|W_n(s_j)|^2}{s_j},$$

where s_0 and s_J are the smallest and largest wavelet scale, respectively.

REFERENCES

Caliński, T., and J. Harabasz, 1974: A dendrite method for cluster analysis. *Commun. Stat.*, **3**, 1–27, doi:10.1080/03610928308827180.

Diabate, L., Ph. Blanc, and L. Wald, 2004: Solar radiation climate in Africa. *Sol. Energy*, **76**, 733–744, doi:10.1016/j.solener.2004.01.002.

Duchon, C. E., and M. S. O'Malley, 1999: Estimating cloud type from pyranometer observation. *J. Appl. Meteor.*, **38**, 132–141, doi:10.1175/1520-0450(1999)038<0132:ECTFPO>2.0.CO;2.

Fovell, R. G., and M. C. Fovell, 1993: Climate zones of the conterminous United States defined using cluster analysis. *J. Climate*, **6**, 2103–2135, doi:10.1175/1520-0442(1993)006<2103:CZOTCU>2.0.CO;2.

JMA, 1996: Synoptic reports at one-minute intervals. Japan Meteorological Agency Business Support Center, CD-ROM.

Lake, D. E., J. S. Richman, M. P. Griffin, and J. R. Moorman, 2002: Sample entropy analysis of neonatal heart rate variability. *Amer. J. Physiol. Regul. Integr. Comp. Physiol.*, **283**, R789–R797, doi:10.1152/ajpregu.00069.2002.

Lave, M., and J. Kleissl, 2010: Solar variability of four sites across the state of Colorado. *Renewable Energy*, **35**, 2867–2873, doi:10.1016/j.renene.2010.05.013.

——, ——, and E. Arias-Castro, 2012: High-frequency irradiance fluctuations and geographic smoothing. *Sol. Energy*, **86**, 2190–2199, doi:10.1016/j.solener.2011.06.031.

Li, S., Q. Zhou, S. Wu, and E. Dai, 2006: Measurement of climate complexity using sample entropy. *Int. J. Climatol.*, **26**, 2131–2139, doi:10.1002/joc.1357.

Marcos, J., L. Marroyo, E. Lorenzo, and M. Garcis, 2012: Smoothing of PV power fluctuations by geographical dispersion. *Prog. Photovolt. Res. Appl.*, **20**, 226–237, doi:10.1002/pip.1127.

Meyers, S. D., B. G. Kelly, and J. J. O'Brien, 1993: An introduction to wavelet analysis in oceanography and meteorology: With application to the dispersion of Yanai waves. *Mon. Wea. Rev.*, **121**, 2858–2866, doi:10.1175/1520-0493(1993)121<2858:AITWAI>2.0.CO;2.

Murata, A., H. Yamaguchi, and K. Otani, 2009: A method of estimating the output fluctuation of many photovoltaic power generation systems dispersed in a wide area. *Electr. Eng. Japan*, **166**, 9–19, doi:10.1002/eej.20723.

Murtagh, F., and P. Legendre, 2014: Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J. Classif.*, **31**, 274–295, doi:10.1007/s00357-014-9161-z.

Ohtake, H., J. G. S. Fonseca Jr., T. Takashima, T. Oozeki, K. Shimose, and Y. Yamada, 2015: Regional and seasonal characteristics of global horizontal irradiance forecasts obtained from the Japan Meteorological Agency mesoscale model. *Sol. Energy*, **116**, 83–99, doi:10.1016/j.solener.2015.03.020.

Pincus, S. M., 1991: Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA*, **88**, 2297–2301, doi:10.1073/pnas.88.6.2297.

——, and A. L. Goldberger, 1994: Physiological time-series analysis: What does regularity quantify? *Amer. J. Physiol. Heart Circ. Physiol.*, **266**, H1643–H1656.

Richman, J. S., and J. R. Moorman, 2000: Physiological time-series analysis using approximate entropy and sample entropy. *Amer. J. Physiol. Heart Circ. Physiol.*, **278**, H2039–H2049. [Available online at http://ajpheart.physiology.org/content/278/6/H2039.]

Tomson, T., and G. Tamm, 2006: Short-term variation of solar radiation. *Sol. Energy*, **80**, 600–606, doi:10.1016/j.solener.2005.03.009.

Torrence, C., and G. P. Compo, 1998: A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.*, **79**, 61–78, doi:10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Science* 3rd ed. Academic Press, 676 pp.

Woyte, A., R. Belmans, and J. Nijs, 2007: Fluctuation in instantaneous clearness index: Analysis and statistics. *Sol. Energy*, **81**, 195–206, doi:10.1016/j.solener.2006.03.001.

Yoshida, S., and H. Kikuchihara, 1989: Solar radiation maps for Japan—Part 10. Climatological divisions for solar irradiance (in Japanese). *J. Japan Sol. Energy Soc.*, **15**, 15–22.

Zagouras, A., A. Kazantzidisa, E. Nikitidoua, and A. A. Argiriou, 2013: Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. *Sol. Energy*, **97**, 1–11, doi:10.1016/j.solener.2013.08.005.

——, R. H. Inman, and C. F. M. Coimbra, 2014a: On the determination of coherent solar microclimates for utility planning and operations. *Sol. Energy*, **102**, 173–188, doi:10.1016/j.solener.2014.01.021.

——, H. T. C. Pedro, and C. F. M. Coimbra, 2014b: Clustering the solar resource for grid management in island mode. *Sol. Energy*, **110**, 507–518, doi:10.1016/j.solener.2014.10.002.