Predictability of Characteristics of Temporal Variation in Surface Solar Irradiance Using Cloud Properties Derived from Satellite Observations

TAKESHI WATANABE AND DAISUKE NOHARA

Central Research Institute of Electric Power Industry, Abiko, Chiba, Japan

(Manuscript received 26 January 2018, in final form 18 August 2018)

ABSTRACT

Understanding of the characteristics of variation in surface solar irradiance on time scales shorter than several hours has been limited because ground-based observation stations are located coarsely. However, satellite observation data can be used to bridge this gap. We propose an approach for predicting characteristics of a time series of surface solar irradiance in a 121-min time window for areas without ground-based measurement systems. Time series features—mean, standard deviation, and sample entropy—are used to represent the characteristics of variation in surface solar irradiance quantitatively. We examine cloud properties over the area to design prediction models of these time series features. Cloud properties averaged over the defined domain and texture features that represent characteristics of the spatial distribution of clouds are used as measures of cloud features. Predictors for time series features, where explanatory variables are cloud features, are constructed employing the random-forest regression method. The performance test for predictions indicates that the mean and standard deviation can be predicted with higher prediction skill, whereas the predictor for sample entropy has lower prediction skill. The importance of cloud features for predictors and partial dependence of the predictors on explanatory variables are also analyzed. Cloud optical thickness (COT) and cloud fraction (CFR) were important for predicting the mean. Two texture features—contrast and local homogeneity (LHM)—and COT were important for predicting the standard deviation, and COT, LHM, and CFR were important for predicting the sample entropy. These results indicate which satellite-derived cloud field properties are useful for predicting time series features of surface solar irradiance.

1. Introduction

Solar power conversion systems are important in addressing global warming. Photovoltaic (PV) electric power systems are a promising technology because of their low lifetime carbon dioxide emissions. The variability of surface solar irradiance destabilizes the power output of PV power plants. This is a major problem with installing and operating PV power plants and will become more serious as more are installed. However, as suggested by [Ela et al. \(2013\)](#), supplying essential reliable information about the variability of surface solar irradiance would reduce instability in the grid system.

We focus on the variation in surface solar irradiance on time scales shorter than several hours. The ground-based observation data for surface solar irradiance are often used to analyze shorter-time-scale variation. Ground-based observations can supply high-frequency

sampling data, although the data represent only a small area around the observation station. Using data derived from satellite observations with a large observation range can overcome this problem.

Previous studies have examined the spatial distribution of the variability. [Lave et al. \(2017\)](#) constructed predictors of high-frequency variability from low-frequency data. The high-frequency variability was evaluated from ground observation data and the low-frequency variability was evaluated from satellite observation data. They demonstrated the solar variability zones over the continental United States and Hawaii. [Watanabe et al. \(2016a\)](#) proposed using cloud feature data derived from satellite observations to predict the variability of surface solar irradiance. They classified the time series of surface solar irradiance according to the variation features. The classifier was constructed based on the relationship between surface solar irradiance variation and cloud features. [Zagouras et al. \(2014\)](#) also used satellite-derived solar irradiance data. They developed the clustering approach to create maps for assessing solar resources. These approaches rely on cloud

Corresponding author: Takeshi Watanabe, watatake@criepi.denken.or.jp

TABLE 1. Definitions of acronyms used in this paper.

Acronym	Definition	Acronym	Definition
ASM	Angular second moment	CSR	Clear-sky restoral
CER	Cloud effective radius	CTH	Cloud-top height at a pressure level
CFR	Cloud fraction	ENT	Entropy
CI	Clearness index	LHM	Local homogeneity
CNT	Contrast	ME	Mean error
COT	Cloud optical thickness	MAE	Mean absolute error
CRR	Correlation	RMSE	Root-mean-square error

information derived from satellite observations and consequently are influenced by the accuracy of satellite observations and the validity of retrieval product data. Thus, it is important to determine the relationship between clouds and surface solar irradiance and to construct a method reflecting this relationship.

In this work, we propose a new approach using cloud properties derived from satellite observations to predict the characteristics of the temporal variation in surface solar irradiance on shorter time scales. Metrics have been introduced to quantify time series characteristics (e.g., [Duchon and O'Malley 1999](#); [Lave and Kleissl 2010](#); [Tomson and Tamm 2006](#); [Watanabe et al. 2016b](#)). It has been suggested that using several metrics is better for representing variation characteristics than using a single metric ([Duchon and O'Malley 1999](#); [Watanabe et al. 2016b](#)). The data processing in this work for representing the variation features follows [Watanabe et al. \(2016a,b\)](#). They used three time series features: mean, standard deviation, and sample entropy (see [section 4](#)). The set of these time series features provides quantitative information about not only the strength of surface solar irradiance but also its variation. Time series features predicted on pixels in the area covered by cloud property data derived from satellite observation enable the spatial distribution of characteristics of the variation to be ascertained. Predictors developed in this work learn the relationship between cloud properties and variation features from training data; thus, we should examine what training data should be used. [Watanabe et al. \(2016a\)](#) suggested that the relationship between surface solar irradiance and cloud properties derived from Moderate Resolution Imaging Spectroradiometer (MODIS) satellite observations is unclear when the surface solar irradiance is moderate. The relationship between surface solar irradiance and cloud properties must be clarified to construct skillful predictors.

The remainder of this paper is organized as follows. [Sections 2 and 3](#) describe the data and methods used in this work. We explain the data processing for the feature extinction of clouds and time series in [section 4](#). In [section 5](#), we investigate the relationship between variation in surface solar irradiance and cloud features, focusing on

the effect of three-dimensional radiative transfer. From the results in previous sections, predictors for time series features are constructed and the performance of predictors is tested in [section 6](#). The results and findings are summarized in [section 7](#). The acronyms used in this paper are listed in [Table 1](#).

2. Data

a. Downward shortwave irradiance at the surface

We use global surface solar irradiance data from Japan ([Japan Meteorological Agency 1996](#)). The Japan Meteorological Agency maintains ground-based observation equipment and performs quality control ([Ohtake et al. 2015](#)). The downward shortwave flux at ground level is provided as 1-min accumulated values with an interval of 1 min. The data are from the 6 years from 2010 to 2015. Surface solar irradiance data from 47 stations are used ([Fig. 1](#)).

The solar irradiance data are normalized by the downward shortwave flux at the top of the atmosphere.

FIG. 1. Locations of ground-based observation stations for surface solar irradiance in Japan.

The normalized surface solar irradiance is called the clearness index (CI) (Woyte et al. 2007):

$$CI(t) = I_s(t)/I_t,$$

where I_s is the surface solar irradiance and I_t is the downward shortwave flux at the top of atmosphere. Flux I_t is calculated considering

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

where the solar constant $I_0 = 1367 \text{ W m}^{-2}$, $E(t)$ is the eccentricity factor at time t , and $Z(t, l)$ is the solar zenith angle at time t and latitude l . As a result of this normalization, the diurnal cycle, seasonal cycle, and effect of latitude on the solar zenith angle are removed. This data processing is convenient for analyzing surface solar irradiance data at different sites in different seasons simultaneously.

b. MODIS cloud properties

We use the level-2 cloud product of MODIS on *Terra* and *Aqua* (Platnick et al. 2015a,b) in product datasets MOD06 and MYD06, respectively. The datasets cover the same period as the surface solar irradiance.

These products contain some physical parameters related to the cloud properties. We use cloud optical thickness (COT), cloud effective radius (CER), cloud-top height at a pressure level (CTH), and cloud fraction (CFR). The resolutions of COT, CER, and CTH are about 1 km, and that of CFR is about 5 km.

Quality information about the algorithm for detecting clouds, called the cloud mask, and retrieval of cloud properties are also stored in the MOD and MYD datasets. We analyze the clear-sky restoral (CSR) flag (Platnick et al. 2014). The MODIS cloud mask is designed to be clear-sky conservative, which means that the cloud detection algorithm of the MODIS data product seeks to identify not cloudy pixels and allows more false clouds than false clear sky (Ackerman et al. 2010). The CSR algorithm is introduced to identify pixels expected to be poor retrieval candidates because of sun glint, edges of clouds, heavy dust or smoke contamination, or spatially variable (partly cloudy) pixels. There are four possible outcomes of the CSR algorithm: overcast cloudy, not cloudy, partly cloudy, and cloud edge. Clear pixels assigned by the MODIS cloud mask are assigned as overcast cloudy by the CSR algorithm. To avoid misunderstandings, the overcast cloudy CSR outcome is called overcast-cloudy/clear in this work.

MOD and MYD datasets store two types of retrieval data in COT and CER. One is the optical property data, where the retrieval is based on the plane-parallel

assumption. The retrieval based on one-dimensional plane-parallel radiative transfer is used for pixels assigned as cloudy and probably cloudy by the MODIS cloud mask algorithm (hereinafter called PP data). The other is retrieval data for pixels that are identified as either partly cloudy or cloud edge by the CSR algorithm. The plane-parallel assumption is not valid for partly cloudy retrieval because the three-dimensional radiative effect becomes significant (Zhang and Platnick 2011). The partly cloudy retrieval is based on the three-dimensional radiative transfer.

The two types of COT and CER data are merged to consider the three-dimensional radiative transfer effects. The value of pixels assigned as partly cloudy and cloud edge by the CSR algorithm in the PP data are replaced with that in the partly cloudy retrieval data. Pixels assigned as not cloudy by the CSR algorithm are reassigned as clear, where COT and CER are undefined. CFR can also be modified considering the CSR algorithm (Platnick et al. 2014). Hereinafter, the merged data are called MRG data. The CSR flag and partly cloudy cloud properties are analyzed in section 4.

Terra and *Aqua* are polar-orbiting satellites. The daytime equatorial crossing times of *Terra* and *Aqua* are 1030 and 1300 Japan standard time, respectively. The ground-based observation stations in Japan are observed when the solar zenith angle is high.

The across-track width in the swash is about 2330 km. The retrieval confidence is decreased farther from the nadir (Maddux et al. 2010). The pixels within the range of one-third of the across-track width from its center are used for analyses.

The MODIS observation time of the pixel over the ground observation station is called the simultaneous observation time in this study. We analyze the surface solar irradiance and cloud properties at the same observation time to combine the two types of data. This is necessary to investigate the relationship between the variation feature of the ground solar irradiance and cloud properties. The sample number of the simultaneous observation time is 58 545 for 6 years.

c. Cloud-motion speed

The Multiangle Imaging SpectroRadiometer (MISR) on the *Terra* satellite measures reflected solar radiation in nine directions distributed along the track. Cloud motion is derived by tracking cloud patterns (Horváth 2013). The level-2 cloud-motion-vector product (Diner 2012) is used from the same period as the MODIS cloud product from 2010 to 2015. We analyze the cloud-motion speed over ground observation stations for surface solar irradiance in Japan.

FIG. 2. Histogram of cloud-motion speed (m s^{-1}). The bin width is 2.5 m s^{-1} .

3. Method

a. Random-forest regression

For the regression analysis, we use the random-forest regression method, which is a method for growing an ensemble of weak tree-based learners (Breiman 2001; Hastie et al. 2009). The random forest builds a large collection of decorrelated trees. Because of this decorrelation, the prediction variance decreases. We use the randomForest package (Liaw and Wiener 2002) in R software (version 3.6; <http://www.R-project.org/>) for the random-forest regression method.

The K -fold cross-validation method is used to verify the prediction skill (Hastie et al. 2009). All data are divided randomly into K parts of similar sizes. A predictor is trained using nine parts, and one remaining part is used as test data. The prediction skill is evaluated employing metrics. This verification test is repeated 10 times for each combination of training and test data. Last, the metrics of the K test results are averaged.

There are two advantages in using the random-forest regression method for this work. One is that we can compute the relative importance of variables, which measures the prediction strength of each explanatory variable (Hastie et al. 2009). The other is that the marginal effect of a variable on the response of the predictor can be estimated with a partial-dependence plot (Friedman 2001). The partial-dependence plot is used to interpret models produced by the black-box prediction method, similar to machine-learning methods. Details about these two approaches are provided in the appendix. The combination of these two random-forest regression-model methods can be interpreted (Hastie et al. 2009). First, the relative importance test is performed to identify the most relevant variables to the regression model. This is because when a large number of input variables is used, only a few variables are affected frequently in the model response. Next, the partial-dependence plot of the chosen variables is

constructed. The curve visualizes the dependence of the model on the variable, allowing us to interpret how each variable affects the model.

b. K -means classification

The K -means classification method is an unsupervised nonhierarchical method (Hartigan and Wong 1979; Hastie et al. 2009). Each point is classified into the most similar (nearest) cluster one by one. Dissimilarity is measured by the Euclidian distance between the point tested and the cluster center. At the end of clustering analysis, the number of groups must be determined. To choose the optimal group number, the K -means classification is repeated with different numbers of groups. In each classification analysis, the separation of clusters and the similarity of members in each group are checked using the Calinski–Harabasz pseudo- F statistic (Calinski and Harabasz 1974). This statistic is defined as

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

where A and W are among- and within-cluster variances, respectively; n is the number of samples; and k is the number of groups.

4. Data processing for feature extinction of clouds and time series

a. Time series features of surface solar irradiance

To quantify the variation features of surface solar irradiance, we use three time series features, namely, the mean, standard deviation, and sample entropy. Three time series matrices are computed for the local time series of the CI in the time window. The mean, standard deviation, and sample entropy represent the magnitude of the CI, the magnitude of its variation, and the complexity of its variation, respectively. When the sample entropy becomes large (small), the time series of solar irradiance tends to fluctuate with higher (lower) frequency (Watanabe et al. 2016b). Using sample entropy

FIG. 3. The *K*-means classification analysis for time series features for (a) four clusters, and (b)–(d) distributions of the time series features for the variation clusters for the seven-cluster case. The colors in (b)–(d) correspond to the definitions of the variation clusters in (a). The black box corresponds to the distribution for all points. The upper and lower box sides are defined as the 75th and 25th percentiles, respectively, and the height of the box corresponds to the interquartile range (IQR). The upper whisker extends to the highest data point at or below the value given by 75th percentile + $(1.5 \times \text{IQR})$. The lower whisker extends to the lowest data point at or above the value given by 25th percentile – $(1.5 \times \text{IQR})$. The filled circles above or below the whiskers represent outliers.

can distinguish CI time series fluctuating with shorter and longer periods.

The time window is set as 121 min, and its center is the simultaneous observation time. The range of the time window is chosen considering that the sample entropy needs at least 100 points to obtain a significant value (Richman and Moorman 2000).

b. Cloud features

The cloud properties are defined over the domain, and the center is at the observation station of the surface solar irradiance. The extent of the domain is $45 \text{ km} \times 45 \text{ km}$. The extent is determined considering the cloud moving distance for the period of the defined time window.

The mean wind speed for all ground-based stations is obtained from the MISR cloud-motion vector (Fig. 2).

The mean cloud-motion speed is about 10.693 m s^{-1} , and the median is 7.970 m s^{-1} . The distribution of wind motion speed is not symmetrical and is skewed toward larger values. The median is used as the representative cloud-motion speed in this work. The cloud moves about 57 km for 121 min. Assuming motion with a constant speed, all of the clouds in a $45 \text{ km} \times 45 \text{ km}$ domain cross above the ground-based observation station in the 121-min time window. Consequently, the surface solar irradiance is affected by the clouds in the domain. We examine how the determination of domain

TABLE 2. The number of members in each cluster (C1–C4) and its ratio against the total number (in parentheses; %).

C1	C2	C3	C4	Total
18 336 (31.32)	17 506 (29.90)	9732 (16.62)	12 971 (22.16)	58 545

FIG. 4. Distributions of the ratio of pixels assigned to each category of CSR flag for variation clusters against the total pixel number in the definition domain for (a) overcast-cloudy/clear, (b) cloud edge, (c) not cloudy, and (d) partly cloudy. Total pixel number in the domain is 2025.

size influences the prediction skill of prediction models in section 6.

The MODIS cloud properties are averaged over the domain. Two means of the MODIS cloud properties are computed. One is the mean over all pixels in the defined domain. The other is the mean over the cloudy pixels in the defined domain. The cloudy pixel means are computed for COT, CER, CTH, and CFR. The overall domain means are computed for two cloud properties, namely, COT and CFR, because the overall domain means of CER and CTH are not informative physically. When the domain is completely cloud free, simultaneous observation points are not used for analysis because the cloud properties cannot be defined.

The texture features represent the characteristics of the spatial distribution of gray tone images (Haralick et al. 1973). Texture features are useful for classifying cloud types (Ameur et al. 2004; Watanabe et al. 2016a). Texture features are computed using a base-10 logarithm of COT in the defined domain. The five texture features are used: angular second moment (ASM), contrast (CNT), correlation (CRR), entropy (ENT), and local homogeneity (LHM). Texture features are chosen following Ameur et al. (2004) and Watanabe et al. (2016a). Logarithmic COT ranges from -2 to 2 because the valid range of COT is from 0.01 to 100.00 . Clear pixels where COT is not defined are assigned as -2 because filling all pixels in the defined domain is necessary to compute texture features.

FIG. 5. Distributions of each cloud feature for variation clusters. DMN and CLD represent the domain- and cloud-averaged values, respectively. Boxplots are drawn in the same manner as in Fig. 3, but outliers are omitted. The units in (c) and (k) are micrometers and hectopascals, respectively; all other variables are unitless.

5. Analysis of the relationship between variation characteristics of the surface solar irradiance and cloud features

a. Classification of simultaneous observation points

All simultaneous observation points are classified using the *K*-means classification method. Each simultaneous point has three time series properties and nine cloud properties. Three-dimensional vectors constructed from the three time series features are used as the feature vectors to measure the dissimilarity. The number of clusters is determined to be four according to PseudoF statistics. Figure 3a shows the cluster analysis results. The resultant clusters are called variation clusters and are labeled from C1 to C4 in ascending order of the mean CI. These results

are similar to Watanabe et al. (2016a), although they analyzed one region in Japan and the number of clusters was six. Table 2 summarizes the number of members in each variation cluster. C1 has the largest number of members of the four clusters. The number of members in C2 and C4 is comparable, although that in C3 is smaller.

The differences in the time series features between variation clusters are shown in Figs. 3b–d. The characteristics of each variation cluster are summarized as follows. C1 is the weakest in terms of mean CI and also shows weak variability. This cluster corresponds to overcast cloudy conditions. C2 and C3 are the moderate mean CI groups. These variation clusters have larger standard deviation, which indicates that their variability is significant. Although these two variation clusters have

FIG. 6. The relationship between the mean of CI and cloud features from MRG data. For clarity, these figures are drawn using 3000 randomly chosen points.

comparable standard deviations, C2 has smaller sample entropy than C3. This difference indicates that the CI time series assigned to C2 fluctuates with a longer period. However, the time series assigned to C3 fluctuates with a shorter period. C4 has the largest mean CI and the variability is weak. This cluster corresponds to the clear and almost clear conditions.

b. Clarification of the relationship between time series features and cloud properties

This section discusses the cloud properties in each variation cluster considering the CSR flag and clarifies the relationship between time series features and cloud properties. Figure 4 shows the distribution of the number of pixels that are assigned as a CSR flag for each CI variation cluster. Most pixels in C1 are identified as overcast-cloudy/clear, although a small number of pixels are assigned as cloud edge. In the other clusters, 75%–95% of all pixels are assigned as overcast-cloudy/clear (Fig. 4a). The second main CSI flag is cloud edge (Fig. 4b). The CSR algorithm defines cloud edge as overcast cloudy pixels with clear adjacent neighbors (Platnick et al. 2014). The domain-averaged CFR of these variation clusters is considerably less than 1.0 (Fig. 5i), which indicates that cloudy and clear pixels coexist in the definition area. The boundary between the

cloud and clear regions can be identified as cloud edge by the CSR algorithm. C3 also has the most partly cloudy pixels. C3 corresponds to a time series with strong, rapid fluctuations and the greatest complexity. The fluctuations are caused by broken and disordered clouds (Martínez-Chico et al. 2011). The large amount of partly cloudy pixels in C3 reflects such cloud properties, although cloud smaller than the MODIS resolution may not be detected accurately. C4 has most pixels assigned as not cloudy (Fig. 4c). However, the number of pixels assigned as not cloudy is small when compared with the two main categories of CSR flag.

We compare the cloud properties of the PP and MRG cloud optical properties data for each variation cluster (Fig. 5). The distributions of the cloud properties cloud-averaged COT and texture features show clear changes after considering the CSR flag. The distributions for C1 barely change because most pixels are assigned as overcast cloudy by the CSR algorithm. However, the distributions of the cloud properties in the other three clusters vary significantly. MRG cloud-averaged COT tends to decrease against PP cloud-averaged COT, although the change in the distribution width is not significant (Fig. 5a). The changes in the texture features are significant. The ranges of distribution in ENT, ASM, and LHM of C2 and

FIG. 7. As in Fig. 6, but for standard deviation.

C3 of the MRG data become narrower than those of the PP data, although the opposite occurs for C4. These results indicate that the relationship between time series features and cloud are clarified by considering the three-dimensional radiative effect. Medians of ENT and CNT of MRG data become greater than those of PP data, and medians of ASM and CRR tend to become smaller. The effect of using MRG data on the prediction skill of the time series features is discussed in the next section.

Figures 6–8 show the relationship between each time series feature and MRG cloud properties. The mean is well explained by COT and CFR because these cloud properties show a unique relationship with the mean (Fig. 6). As COT increases, the mean decreases, and CFR shows the same behavior. This relationship is exponential-like. In contrast, the standard deviation is not as clearly explained by specific cloud properties (Fig. 7). For example, there is a positive correlation between the standard deviation and CNT. CNT increases with standard deviation; however, the variance is large. It is difficult to explain the relationship between sample entropy and cloud features based on Fig. 8. We discuss these relationships based on the relative importance analysis using the random-forest regression in the next section.

CTH is often used to determine cloud type and cloud phase (e.g., Platnick et al. 2014). The CTH distribution

differs between variation clusters (Fig. 5k). CTH in C1 tends to be higher than 600 hPa and lower than 400 hPa in C2–C4. The CTH distribution in every cluster is wide, indicating that there are several cloud types according to CTH in each variation cluster. Figures 6c, 7c, and 8c show there are two major groups, lower and higher CTH, and that relationships between each time series feature and CTH are not clear.

6. Prediction of time series features

The relationship between the time series features and cloud properties is not simple, and predictors are needed to learn the complex systems. The random-forest regression method is used to construct predictors for time series features. MRG cloud features are used as their explanatory variables. The training data are the simultaneous observation dataset of *Terra* and *Aqua* from 2010 to 2015. The number of ensemble trees is 500. As the number of trees increases, the error variance gradually decreases; 500 trees are enough to stabilize the error variance. Predictors are valid only when there are clouds in the domain and the domain that is completely cloud-free is not the target of the prediction. This is because cloud properties are not defined in the cloud-free domain, as mentioned in section 4b. The relationships between the time series features and cloud properties are

FIG. 8. As in Fig. 6, but for sample entropy.

based on the simultaneous observation data. Thus, the prediction models do not explicitly include several effects related to cloud, such as cloud motion and cloud birth and death. Consequently, these effects are most likely to degrade the prediction skill in the whole range of each time series feature.

To estimate the performance of predictors, K -fold cross validation is performed. All data are divided into 10 parts, and thus $K = 10$. The validation process is repeated 10 times. The metrics for the performance of predictors are the mean error (ME), the mean absolute error (MAE), and the root-mean-square error (RMSE). In addition, standardized metrics, which are the three metrics divided by the mean of each time series feature, are used (%ME, %MAE, and %RMSE). After repeating the validation procedures 10 times, each of the metrics is averaged.

The results of the performance test are shown in Fig. 9 and Table 3. The predictor for the mean shows better prediction skill. The relatively large error is seen in the large value, where most points correspond to C4. The standard deviation predictions also correspond well to the observations, but the prediction error becomes large when the observed standard deviation is greater than 0.15. The major cluster included in this large error area is C2 and C3. The large MAE and RMSE are due to this prediction error. The predictor for the standard

deviation tends to predict larger (smaller) values in the small (large) standard deviation region. For the sample entropy, the predictor's skill is lower than for the other two time series features. Prediction skill decreases considerably when the observations are larger than 0.5. Most samples in these low prediction areas belong to C3 (Figs. 9e,f).

The relative importance of cloud properties for each predictor is shown in Fig. 10. The partial dependences of the three most important variables on each time series feature are shown in Fig. 11. Important variables for the mean predictor are domain-averaged COT and domain- and cloud-averaged CFR. As domain-averaged COT and domain-averaged CFR increase, the predicted mean decreases. These results are reasonable when the mean CI in the time window is considered because the domain-averaged COT and CFR are related by the decay of the downward shortwave irradiance. Figure 10a shows that the relative importance of the domain-averaged CFR and cloud-averaged CFR for the predictor of the mean are similar. In addition, the shapes of the partial-dependence curves of the two CFRs are similar (Figs. 11b,c). Thus, these two CFRs must be collinear for the predictor of the mean. The important variables for the standard deviation predictors are domain-averaged CFR and the two texture features, CNT and LHM. The differences between these three cloud properties are small. The dependence curve

FIG. 9. Estimation of the prediction skill for each predictor: shown are scatterplots (left) between observations and predictions and (right) between observations and prediction errors (prediction minus observation) for the (a),(b) mean, (c),(d) standard deviation, and (e),(f) sample entropy. The boxplots in (b), (d), and (f) show the distribution of prediction error in the observation bin. There are 10 bins at equal observation intervals. Boxplots are drawn in the same manner as in Fig. 3. The colors correspond to the variation cluster for the seven-cluster case in Fig. 3b.

TABLE 3. Metrics (ME, MAE, and RMSE; unitless) used to estimate prediction skill for each time series feature predictor, along with the metrics standardized by the mean of observation of each time series feature (in parentheses; %).

Time series feature	Mean of obs	ME	MAE	RMSE
Mean	0.4338	−0.0002 (−0.05)	0.0598 (13.78)	0.0825 (19.02)
Std dev	0.1083	0.0007 (0.61)	0.0366 (33.83)	0.0488 (45.06)
Sample entropy	0.2658	0.0081 (3.05)	0.1328 (49.97)	0.1988 (74.81)

of domain-averaged CFR has a peak at around 0.6. CNT has a positive relationship with standard deviation; however, LMH has a negative relationship. Considering CNT represents the local variation in the image, a large CNT corresponds to an image where COT changes spatially with large differences between neighboring pixels. LHM is the opposite of CNT and represents the homogeneity of COT between neighboring pixels. There may be negative collinearity between CNT and LHM for the standard deviation predictor. The results for standard deviation show that the amplitude of variation in the surface solar irradiance results from cloudy areas where the thick and thin COT coexist or cloudy and clear pixels coexist. Domain-averaged COT, LHM, and

domain-averaged CFR are the most important variables in descending order for the predictor for sample entropy. The relative importance of domain-averaged COT overwhelms the other two important cloud properties. Large sample entropy is observed in the cloudy area where domain-averaged COT is small and CFR is large, but it is not perfectly overcast. Additionally, clouds in the defined domain are not homogenous.

The differences in prediction skill between using MRG and PP data are summarized in Table 4. These results show the benefits of using the MRG data. For all of the time series features, all the prediction skill metrics are improved by using MRG data. Comparing standardized metrics, the improvement in the mean is the

FIG. 10. Relative importance of cloud properties for each of the three predictors: (a) mean, (b) standard deviation, and (c) sample entropy. The black bars indicate the top three cloud properties for each predictor.

FIG. 11. Partial dependence of each predictor on the three most important variables: (a)–(c) mean, (d)–(f) standard deviation, and (g)–(i) sample entropy predictors. The important variable is on the horizontal axes, and partial dependence is on the vertical axes.

smallest of the three time series features, and the prediction skill of the standard deviation and sample entropy show comparable improvements. The effects of using MRG data on texture features are considerable, as discussed in section 5 (Fig. 5). These relative importance analysis results show that LHM is important for predicting both the standard deviation and sample entropy. The change in texture features due to the MRG data likely contributes to the improvement of the prediction skill.

Although CTH is often used to determine cloud type and cloud phase, CTH is not in the top three relative importance variables for any predictors. This result indicates that cloud types classified based on CTH are not strongly related to the variation characteristics of the surface solar irradiance.

To examine the sensitivity of the prediction models to domain size, the predictive capabilities of the models are compared under various domain sizes. In all cases, the time window of the time series is 121 min. Three domain

TABLE 4. Difference (defined as the value of a metric in the MRG case minus that in the PP case) in metrics (ME, MAE, and RMSE; unitless) used to estimate prediction skill for each time series feature predictor, along with the difference in metrics standardized by the mean of observation of each time series feature (in parentheses; %).

Time series feature	ME	MAE	RMSE
Mean	−0.0001 (−0.02)	−0.0003 (−0.06)	−0.0003 (−0.63)
Std dev	−0.0001 (−0.05)	−0.0009 (−0.85)	−0.0011 (−1.00)
Sample entropy	−0.0003 (−0.12)	−0.0024 (−0.83)	−0.0028 (−0.97)

sizes are compared: 25 km × 25 km, 45 km × 45 km, and 115 km × 115 km (see Table 5). The middle case, 45 km × 45 km, best predicts the mean. The cases with 25 km × 25 km and 45 km × 45 km are similar on prediction of standard deviation and sample entropy, and the 115 km × 115 km case is worse. These results indicate that the model's predictive ability is sensitive to the defined domain of the cloud properties and that information about cloud motion is necessary to clarify the relationship between time series features and cloud properties.

Last, we show the prediction of three time series features on pixels of MODIS cloud properties (Fig. 12). From a set of these figures, we obtain the spatial distribution of surface solar irradiance variation. Larger standard deviation is seen at the boundary between the low and middle mean CI and between the cloudless and cloudy zone. Large sample entropy is also seen at the cloudless–cloudy boundary, which indicates that the surface solar irradiance fluctuates with a shorter period and larger amplitude.

7. Summary and discussion

The understanding of the spatial distribution of the variation features of surface solar irradiance has been inadequate because of the limited ground-based observations. Therefore, we developed an approach using cloud properties derived from satellite observations to predict the time series features of the surface solar irradiance. The predictors were constructed based on the relationship between the cloud properties and variation features of the surface solar irradiance. We investigated this relationship.

Three time series features, namely, mean, standard deviation, and sample entropy, were used to represent

the variation features of the surface solar irradiance. Cloud features were obtained from domain- and cloud-averaged cloud properties and texture features, which were used to represent the characteristics of spatial distribution of cloud. We analyzed the ground-observed global solar irradiance data and cloud product derived from simultaneous MODIS observations. The MODIS cloud product contains two types of retrieved optical properties data. One is based on the plane-parallel assumption, and the other is based on three-dimensional radiative transfer. The two types of MODIS cloud optical properties data were merged to consider the three-dimensional radiative effect. The merged data clarified the relationship between time series features and cloud properties. The effect of three-dimensional radiative transfer is expected to be significant when there is coverage with low and intermediate COT. The CSR flag analysis indicated that surface solar irradiance fluctuations with a high amplitude, which corresponded to a large standard deviation, were related to partly cloudy retrieval data. In this case, the effect of three-dimensional radiative transfer was large.

Predictors for time series features were constructed to apply the random-forest regression method. The merged cloud properties were used as explanatory variables. The predictors for mean and standard deviation had better prediction skill, although the predictor for sample entropy had a lower prediction skill. The importance and partial dependence of explanatory variables for the predictor were analyzed. COT and CFR were important for predicting the mean. CNT, LHM, and COT were important for standard deviation, and COT, LHM, and CFR were important for sample entropy. These results indicate that the retrieved cloud

TABLE 5. Comparison of prediction skill for different domain sizes. The values shown are the metrics (ME, MAE, and RMSE) standardized by the mean of observation of each time series feature (%), and those given for the 45 km × 45 km case are the same as in Table 3.

Time series feature	%ME			%MAE			%RMSE		
	25 km × 25 km	45 km × 45 km	115 km × 115 km	25 km × 25 km	45 km × 45 km	115 km × 115 km	25 km × 25 km	45 km × 45 km	115 km × 115 km
Mean	−0.07	−0.05	−0.16	14.92	13.78	15.60	19.96	19.02	21.32
Std dev	0.57	0.61	0.77	33.43	33.83	39.69	43.72	45.06	51.87
Sample entropy	2.90	3.05	3.50	50.15	49.97	56.69	73.47	74.81	83.46

FIG. 12. Predictions for each time series feature for the MODIS field-of-view pixels: (a) mean, (b) standard deviation, and (c) sample entropy. The cloud product used is MYD06.A2014009.0345. The pixels within the range of one-third of the across-track width from the center are used for analyses. Prediction is not performed on the gray pixels because either they are part of a cloud-free domain or they are outside the analysis region.

properties and the spatial distribution of clouds affected the variation features.

We used the MODIS cloud product in this work because this cloud product provides useful, accurate information about clouds. However, our approach using the MODIS cloud product is not suitable for near-real-time prediction for application to solar energy engineering because *Terra* and *Aqua* are polar-orbit satellites. Using data derived from geostationary satellites is better for solar energy engineering. As shown in this work, COT, CFR, and texture features, which are calculated from COT, are important for predicting variation features of surface solar irradiance. These cloud features are obtained from new geostationary satellites. For example, cloud optical properties can be retrieved from *Himawari-8* and *Himawari-9* observations (Bessho et al. 2016). Therefore, we expect that our approach can be applied to the *Himawari-8* and *Himawari-9* observations, although the prediction skill must be checked because of differences in retrieval error and spatial resolution between the Advanced Himawari Imager sensors on Himawari and MODIS.

There are various methods for estimating or retrieving solar radiation from databased on satellite observations (e.g., Ineichen and Perez 1999; Takenaka et al. 2011; Xie et al. 2016). These methods provide snapshot images of the surface solar irradiance. To obtain a time series or trends for the surface solar irradiance, satellite observation data at each observation time must be processed, resulting in high computational cost and a delay in providing the data because of the processing time. Our method provides information about the mean and variability of the surface solar irradiance for 121 min. Our prediction method has several important advantages compared with previous work. One of these advantages is the reduction of the processing frequency. Assuming

the atmospheric conditions do not change greatly over 121 min, the prediction process is done every 2 h. In practice, the prediction frequency should be determined based on the atmospheric conditions. Another advantage is that the delay due to the processing time is much shorter for our method because the set of predicted time series features summarizes the characteristics of surface solar irradiance over the 121-min period.

Our proposed prediction method is not limited to the three time series features and the 121-min time window. Our method can be applied to other time series features that evaluate different characteristics of CI time series from those used in this work, and for time series in a time window shorter or longer than 121 min. We discuss two concerns about the choice of the time window here. First, as the time window shortens, the confidence and robustness of the time series features may be reduced; for example, time series data composed of more than 100 points are necessary to obtain reliable sample entropy (Richman and Moorman 2000). We assume that the atmospheric conditions and cloud properties do not change in 121 min. As the time window lengthens, this assumption may not be valid. This causes an unclear relationship between time series features and cloud properties, which are derived from one satellite image, and thus decreases the prediction skill. The second concern is about the resolution of the satellite sensor. The characteristics of the time series of the surface solar irradiance in a shorter time window are related to cloud on a small spatial scale more than on a large spatial scale. If the resolution is low when compared with the cloud size related to the variation scale of the surface solar irradiance, the predictor would not learn the relationship between time series features and cloud properties well.

Acknowledgments. The *Terra* and *Aqua* MODIS level-2 cloud-product datasets were acquired from the Level-1 and Atmosphere Archive and Distribution System (LAADS) Distributed Active Archive Center (DAAC), located in the Goddard Space Flight Center in Greenbelt, Maryland (<https://ladsweb.nascom.nasa.gov/>). The MISR cloud-motion product was obtained from the NASA Langley Research Center Atmospheric Science Data Center (<https://eosweb.larc.nasa.gov/>).

APPENDIX

Relative Importance of Variables and the Partial-Dependence Plot

Relative importance $I_l^2(T)$ for each predictor variable X_l for a single decision tree T is measured by using

$$I_l^2(T) = \sum_{t=1}^{J-1} i_t^2 I[v(t) = l],$$

where $v(t)$ is the splitting variable associated with node t and i_t^2 is the corresponding empirical improvement as a result of the split. In this work, the improvement is defined as the decrease in node impurities from splitting on the variable, and the node impurity is measured by using the residual sum of squares after the split at node t .

For the ensemble tree-based method, relative importance is obtained by averaging $I_l^2(T_m)$ over all of the trees:

$$I_j^2 = \frac{1}{M} \sum_{m=1}^M I_l^2(T_m),$$

where M is the number of ensembles.

A partial-dependence plot is the plot of the partial dependence of a function on the variables of interest. The regression model F predicts the objective variable y as a function of p explanatory variables $X = (x_1, \dots, x_i, \dots, x_p)$. The training data include N observations: $x_{i,j}$ and y_j , $j = 1, 2, \dots, N$. The regression model can be written in the following form:

$$\hat{y}_k = F(X_k) = F(x_{1,k}, x_{2,k}, \dots, x_{p,k}).$$

The partial-dependence function is defined as

$$G(x) = \frac{1}{N} \sum_{k=1}^N F(x_{1,k}, \dots, x_{j-1,k}, x, x_{j+1,k}, \dots, x_{p,k}).$$

Here, G represents the effect of variables x on $F(X)$ after accounting for the average effects of the other explanatory variables.

REFERENCES

- Ackerman, S., R. Frey, K. Strabala, Y. Liu, L. Gumley, B. Baum, and P. Menzel, 2010: Discriminating clear-sky from cloud with MODIS. Algorithm Theoretical Basis Document (MOD35), University of Wisconsin–Madison Cooperative Institute for Meteorological Satellite Studies Doc. (version 6.1), 117 pp.
- Ameur, Z., S. Ameur, A. Adane, H. Sauvageot, and K. Bara, 2004: Cloud classification using the textural features of Meteosat images. *Int. J. Remote Sens.*, **25**, 4491–4503, <https://doi.org/10.1080/01431160410001735120>.
- Bessho, K., and Coauthors, 2016: An introduction to *Himawari-8/9*—Japan's new-generation geostationary meteorological satellites. *J. Meteor. Soc. Japan*, **94**, 151–183, <https://doi.org/10.2151/jmsj.2016-009>.
- Breiman, L., 2001: Random forest. *Mach. Learn.*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.
- Calinski, T., and J. Harabasz, 1974: A dendrite method for cluster analysis. *Commun. Stat.*, **3**, 1–27, <https://doi.org/10.1080/03610927408827101>.
- Diner, D., 2012: MISR level 2 cloud heights and winds HDF-EOS file—Version 1. NASA Langley Research Center Atmospheric Science Data Center DAAC, accessed 24 November 2017, https://doi.org/10.5067/terra/misr/ml2tcsp_j2.001.
- Duchon, C. E. and M. S. O'Malley, 1999: Estimating cloud type from pyranometer observations. *J. Appl. Meteor.*, **38**, 132–141, [https://doi.org/10.1175/1520-0450\(1999\)038<0132:ECTFPO>2.0.CO;2](https://doi.org/10.1175/1520-0450(1999)038<0132:ECTFPO>2.0.CO;2).
- Ela, E., V. Diakov, E. Ibanez, and M. Heaney, 2013: Impacts of variability and uncertainty in solar photovoltaic generation at multiple timescales. National Renewable Energy Laboratory Tech. Rep. NREL/TP-5500-58274, 34 pp., <https://www.nrel.gov/docs/fy13osti/58274.pdf>.
- Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232, <https://doi.org/10.1214/aos/1013203451>.
- Haralick, R. M., K. Shunmugam, and I. Dinstein, 1973: Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, **SMC-3**, 610–621, <https://doi.org/10.1109/TSMC.1973.4309314>.
- Hartigan, J. A., and M. A. Wong, 1979: A K -means clustering algorithm. *Appl. Stat.*, **28**, 100–108, <https://doi.org/10.2307/2346830>.
- Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning*. 2nd ed. Springer, 745 pp.
- Horváth, Á., 2013: Improvements to MISR stereo motion vectors. *J. Geophys. Res. Atmos.*, **118**, 5600–5620, <https://doi.org/10.1002/jgrd.50466>.
- Ineichen, P., and R. Perez, 1999: Derivation of cloud index from geostationary satellites and application to the production of solar irradiance and daylight illuminance data. *Theor. Appl. Climatol.*, **64**, 119–130, <https://doi.org/10.1007/s007040050116>.
- Japan Meteorological Agency, 1996: Synoptic reports at one-minute intervals. Japan Meteorological Business Support Center, CD-ROM.
- Lave, M., and J. Kleissl, 2010: Solar variability of four sites across the state of Colorado. *Renewable Energy*, **35**, 2867–2873, <https://doi.org/10.1016/j.renene.2010.05.013>.
- , R. J. Broerick, and M. J. Reno, 2017: Solar variability zone: Satellite-derived zones that represent high-frequency ground variability. *Sol. Energy*, **151**, 119–128, <https://doi.org/10.1016/j.solener.2017.05.005>.
- Liaw, A., and M. Wiener, 2002: Classification and regression by randomForest. R News, No. 2(3), R Foundation, Vienna,

- Austria, 18–22, https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf.
- Maddux, B. C., S. A. Ackerman, and S. Platnick, 2010: Viewing geometry dependencies in MODIS cloud product. *J. Atmos. Oceanic Technol.*, **27**, 1519–1528, <https://doi.org/10.1175/2010JTECHA1432.1>.
- Martínez-Chico, M., F. J. Batlles, and J. L. Bosch, 2011: Cloud classification in a Mediterranean location using radiation data and sky images. *Energy*, **36**, 4055–4062, <https://doi.org/10.1016/j.energy.2011.04.043>.
- 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, <https://doi.org/10.1016/j.solener.2015.03.020>.
- Platnick, S., and Coauthors, 2014: MODIS cloud optical properties: User guide for collection 6 level-2 MOD06/MYD06 product and associated level-3 datasets. NASA Goddard Space Flight Center Doc., 141 pp., https://modis-images.gsfc.nasa.gov/_docs/C6MOD06OPUserGuide.pdf.
- , and Coauthors, 2015a: *Terra* MODIS Atmosphere L2 Cloud Product (06_L2). NASA MODIS Adaptive Processing System, NASA Goddard Space Flight Center, accessed 24 November 2017, https://doi.org/10.5067/MODIS/MOD06_L2.006.
- , and Coauthors, 2015b: *Aqua* MODIS Atmosphere L2 Cloud Product (06_L2). NASA MODIS Adaptive Processing System, NASA Goddard Space Flight Center, accessed 24 November 2017, https://doi.org/10.5067/MODIS/MYD06_L2.006.
- 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, <https://doi.org/10.1152/ajpheart.2000.278.6.H2039>.
- Takenaka, H., T. Y. Nakajima, A. Higurashi, A. Higuchi, T. Takamura, R. T. Pinker, and T. Nakajima, 2011: Estimation of solar radiation using a neural network based on radiative transfer. *J. Geophys. Res.*, **116**, D08215, <https://doi.org/10.1029/2009JD013337>.
- Tomson, T., and G. Tamm, 2006: Short-term variation of solar radiation. *Sol. Energy*, **80**, 600–606, <https://doi.org/10.1016/j.solener.2005.03.009>.
- Watanabe, T., Y. Oishi, and T. Nakajima, 2016a: Characterization of surface solar-irradiance variability using cloud properties based on satellite observations. *Sol. Energy*, **140**, 83–92, <https://doi.org/10.1016/j.solener.2016.10.049>.
- , T. Takamatsu, and T. Nakajima, 2016b: Evaluation of variation in surface solar irradiance and clustering of observation stations in Japan. *J. Appl. Meteor. Climatol.*, **55**, 2165–2180, <https://doi.org/10.1175/JAMC-D-15-0227.1>.
- Woyte, A., R. Belmans, and J. Nijs, 2007: Fluctuation in instantaneous clearness index: Analysis and statistics. *Sol. Energy*, **81**, 195–206, <https://doi.org/10.1016/j.solener.2006.03.001>.
- Xie, Y., M. Sengupta, and J. Dudhia, 2016: Fast all-sky radiation model for solar applications (FARMS): Algorithm and performance evaluation. *Sol. Energy*, **135**, 435–445, <https://doi.org/10.1016/j.solener.2016.06.003>.
- Zagouras, A., H. T. C. Pedro, and C. F. M. Coimbra, 2014: Clustering the solar resource for grid management in island mode. *Sol. Energy*, **110**, 507–518, <https://doi.org/10.1016/j.solener.2014.10.002>.
- Zhang, Z., and S. Platnick, 2011: An assessment of differences between cloud effective particle radius retrievals for marine water clouds from three MODIS spectral bands. *J. Geophys. Res.*, **116**, D20215, <https://doi.org/10.1029/2011JD016216>.