



A model tree approach to forecasting solar irradiance variability

T.C. McCandless^{a,b,*}, S.E. Haupt^{a,b}, G.S. Young^b

^a National Center for Atmospheric Research, 3450 Mitchell Lane, Boulder, CO 80301, United States

^b The Pennsylvania State University, Department of Meteorology, 503 Walker Building, University Park, PA 16802-5013, United States

Received 12 November 2014; received in revised form 2 June 2015; accepted 14 July 2015

Available online 21 August 2015

Communicated by: Associate Editor Christian A. Gueymard

Abstract

As the penetration of solar power increases, the variable generation from this renewable resource will necessitate solar irradiance forecasts for utility companies to balance the energy grid. In this study, the temporal irradiance variability is calculated by the temporal standard deviation of the Global Horizontal Irradiance (GHI) at eight sites in the Sacramento Valley and the spatial irradiance variability is quantified by the standard deviation across those same sites. Our proposed artificial intelligence forecasting technique is a model tree with a nearest neighbor option to predict the irradiance variability directly. The model tree technique reduces the mean absolute error of the variability prediction between 10% and 55% compared to using climatological average values of the temporal and spatial GHI standard deviation. These forecasts are made at 15-min intervals out to 180-min. A data denial experiment showed that the addition of surface weather observations improved the forecasting skill of the model tree by approximately 10%. These results indicate that the model tree technique can be implemented in real-time to produce solar variability forecasts to aid utility companies in energy grid management.

© 2015 Elsevier Ltd. All rights reserved.

Keywords: Solar irradiance; Model tree; Artificial intelligence; Solar power prediction; Irradiance variability

1. Introduction

The finite supply of fossil fuels and their contribution to climate change will require an increasing utilization of renewable energy. A major concern is that the power generation is variable from renewable energy sources like wind and solar. Solar energy, in particular, varies over a broad range of space and time scales because of the complex dynamic evolution of clouds. Lew et al. (2012) provided evidence of the challenge of solar power integration with

the results showing that the variability of power output was higher with high penetrations of solar than with high penetrations of wind. An important issue with generation of power from solar energy is the lack of storage for photovoltaic (PV) systems, meaning that the energy must be used as it is produced and cannot be effectively stored under current technology constraints. Concentrated Solar Power (CSP) systems, on the other hand, use mirrors or lenses to concentrate sunlight, often used to heat a substance (liquids, salts, etc.) in order to thermally store energy for some period of time, damping the power fluctuations.

Rapid changes in weather conditions, especially cloud growth, advection, and dissipation, cause variability in solar power, particularly from PV systems. Because sparse meteorological data do not fully resolve the cloud field,

* Corresponding author at: National Center for Atmospheric Research, 3450 Mitchell Lane, Boulder, CO 80301, United States. Tel.: +1 303 497 8700.

E-mail addresses: mccandle@ucar.edu (T.C. McCandless), haupt@ucar.edu (S.E. Haupt), g3y@psu.edu (G.S. Young).

deterministic forecasts do not capture all of this variability. Therefore, direct forecasts of solar temporal and spatial variability from the available weather data is valuable for electricity production and transmission decision-makers to manage the power grid efficiently.

In an effort to maximize solar PV power as an energy resource, utility companies require back-up energy sources to balance the power supply with the expected demand. Any difference between the supply and demand is made up with a utility company's operating reserves. The response speed (ramp rate and start time), response duration, frequency of use (continuously or only during rare events), direction of use (up or down), and type of control characterize these operating reserves (Ela et al., 2013). This balancing becomes more challenging when the energy resource is variable. Curtright and Apt (2008) have shown that the cost of energy can be strategically minimized with knowledge of the short and long term PV variations; thus, accurate solar power forecasts provide information to balance operating reserves.

While direct forecasting of spatial and temporal variability of solar irradiance is relatively new, there have been numerous studies aimed at providing deterministic forecasts of the expected value of solar irradiance. Because the time and space dependent field of solar irradiance results from the complex evolution of clouds in the atmosphere, many of these studies have tested non-linear methods for deterministic solar irradiance prediction (Mellit, 2008; Martin et al., 2010; Pedro and Coimbra, 2012; Bhardwaj et al., 2013; Bouzerdoum et al., 2013; Diagne et al., 2013; Fu and Cheng, 2013; Marquez et al., 2011; Inman et al., 2013; Fernandez et al., 2014; Chu et al., 2013; Almonacid et al., 2014; Quesada-Ruiz et al., 2015, among others). While successful, these deterministic forecasts cannot directly capture all of the spatial and temporal variability of solar irradiance, because the available weather data does not resolve all of the cloud scales.

While offering the potential to resolve this problem, direct forecasts of the spatial and temporal variability of the irradiance has been limited. Lave et al. (2013) used a wavelet-based model to predict a power plant's output given a spatio-temporal correlation function and to estimate the variability ratio over the plant's spatial coverage. A more easily generalized approach based on correlation analysis was used by Hoff and Perez (2012) to predict the short-term maximum power output variability based on satellite derived irradiances. This current study addresses this problem using the model-tree non-linear forecasting method (Witten and Frank, 2005).

The temporal scale over which to compute variability depends on the needs of the electric grid management entities. United States utility companies are concerned with 15–30 min fluctuations that achieve power balance with stand-by generators or storage/load management for short range solar irradiance predictions for large cities and dense transmission networks (Perez et al., *in press*). Specifically, the Sacramento Municipal Utility District (SMUD) is

concerned with maximum and minimum values over a 5-min interval (Bartholomy et al., 2014). We therefore use 15 min windows in order to meet the needs of SMUD while still providing an adequate sample size for computation of our variability metric, standard deviation (see Section 2).

The quantification of temporal solar irradiance variability caused by the dynamic evolution of clouds has been extensively studied. Mills et al. (2009) showed a passing cloud at a point produces solar insolation variation exceeding 60% of peak insolation in a matter of seconds. Hinkelman et al. (2007) found that not only are the irradiances themselves larger in the middle of the day but also the fractional change in irradiance from one time to another is larger. Hinkelman (2013) also determined that cloud optical depth and cloud height are the best predictors of irradiance variability at 1 min time resolution. Kuszmaul et al. (2010) analyzed 1-s PV output data and showed that it is linearly proportional to the spatial average of irradiance. Reikard (2009) examined data at resolutions of 5, 15, 30, and 60 min and found that the “data exhibits nonlinear variability, due to variations in weather and cloud cover.” These studies have examined the variability of measured solar irradiance due to changes in cloud cover.

In addition to temporal variability, several studies have examined the spatial variability of solar irradiance. Zagouras et al. (2014) used cluster analysis to determine coherent zones of Global Horizontal Irradiance for utility scale territory in California and used step changes of the daily average clear-sky index at each location to characterize the fluctuation of GHI. Gueymard and Wilcox (2011) analyzed solar power's regional dependence and showed greater variability tends to occur in coastal areas, particularly the California coast, and mountainous areas because of the micro-climate effects of topography. Rayl et al. (2013) performed an irradiance co-spectrum analysis and concluded that solar power site aggregation could greatly reduce power variability on short time scales depending on the distance between sites.

The goal of our study is to use observed meteorological data together with a network of irradiance observing sites to predict both the temporal variability and the spatial variability, both measured in terms of a standard deviation from the mean value of the GHI, and to test those prediction methodologies within the Sacramento, California region. The focus is on short range predictions, which as Nguyen and Kleissl (2014) state, “intra-hour solar forecasting for power production and ramp events has become an important need in the solar industry as the inevitable variability of solar power will have a greater impact on energy resource management as solar penetration increases.” Short range predictions are defined here as the forecasts for up to three hours lead-time provided at 15-min intervals.

An artificial intelligence technique called a model tree, or Cubist model, is used to predict the temporal and spatial

variability, here defined as the standard deviation, of the GHI. While this same approach could also be used to make a deterministic prediction of the expected value of solar irradiance, we focus exclusively on variability in this study. Parallel work on the deterministic forecast problem is presented in [McCandless et al. \(2015\)](#). Section 2 describes the observational data and Section 3 explains the forecasting techniques. The results for temporal and spatial standard deviation predictions are presented in Section 4 while Section 5 illustrates the improvements of the model tree output by inclusion of surface weather observations. The application of the model tree predictions to four specific days is described in Section 6 while the final Section 7 provides discussions and conclusions.

2. Data

2.1. Irradiance observation network

The irradiance observation network used in this study is that of the Sacramento Municipal Utility District (SMUD) in Sacramento, California. We consider the data from the eight solar power forecast sites provided in the SMUD dataset, which are plotted in [Fig. 1](#) as green stars. The eight utility scale PV array sites that measure GHI are geographically diverse across the SMUD service region, which is approximately 900 square miles and includes nearly 100 MW of solar power PV capacity ([Bartholomy et al., 2014](#)). The GHI observations are available from January 25th, 2014 through May 27th, 2014 for a total of 122 days. The resolution of the raw data is 1 min and the solar variability is computed as a standard deviation from the mean value predicted over each 15 min interval. This interval length was selected after communication with several US utility companies and agrees with the shortest time range for which a forecast is useable to the utility. Defining the solar variability as the standard deviation over a 15-min interval means that both the diurnal change in irradiance and the cloud cover resulting change in irradiance are captured in the irradiance variability calculated. This total variability in a 15-min period is the variable most important for utility companies to plan how to allocate the energy reserves and balances this need with the need to know the 5-min maximum and minimum irradiance estimates. The 15-min temporal GHI standard deviation is calculated independently for each SMUD site and only daylight hours are retained in the dataset and is defined as:

$$STDEV_{\text{temporal}}(t) = \sqrt{\frac{\sum_{t_i} (GHI(t_i) - GHI_{t_mean})^2}{n}}, \quad (1)$$

where the GHI_{t_mean} is an average over the 15-min interval, t_i is the 1-min resolution time-step from $t-15$ min to t , and n is the number of times (15).

To compute the spatial variability across the eight SMUD sites, we first compute the 15-min average of the

GHI for each site. Then, the standard deviation of the 15-min GHI averages across the eight sites is calculated as

$$STDEV_{\text{spatial}}(t) = \sqrt{\frac{\sum_{x_i} (GHI(x_i) - GHI_{s_mean})^2}{n}}, \quad (2)$$

where GHI_{s_mean} is the average for all SMUD sites at the time t , x is the notation for a SMUD site, and n is the total number of SMUD sites (8).

Only instances when the observed 15-min average GHI is greater than 20 W/m² are included in the dataset because the goal is to predict only during the daylight hours. Hereafter, the GHI interval is designated by the ending time; thus, a GHI temporal standard deviation prediction at 180 min is actually the temporal standard deviation for the time period from 165 to 180 min.

2.2. Meteorological observation network

It is important to augment the utility irradiance observation network with the nearest meteorological observations. The weather observation network used here is the Meteorological Aviation Report (METAR) network, which are hourly surface weather observation stations typically located at airports across the United States. The METAR observations are quality controlled and processed for ingestion into the National Center for Atmospheric Research (NCAR) Dynamic Integrated foreCast (DICAst) System ([Mahoney et al., 2012](#)). The closest METAR sites to the Sacramento area are the four locations plotted as red X's in [Fig. 1](#). Twelve weather variables are either recorded directly by the METAR stations or are derived as probability values by the NCAR processing system: cloud cover, dewpoint temperature, probability of fog, probability of precipitation in the last hour, probability of precipitation in the last three hours, quantitative precipitation in the last hour, quantitative precipitation in the last three hours, temperature, visibility, wind speed, north-south wind component, and east-west wind component. Therefore, the entire weather observation predictor dataset includes 12 observed weather variables for each hour at four stations for a total of 48 observations.

2.3. Training data

To create a dataset of SMUD GHI observations matched with the hourly METAR weather observations, each 15-min SMUD GHI interval is matched with the corresponding METAR observations at the top of the next hour. For example, the irradiance observation at 1/26/2014 15:15, 15:30, 15:45, and 16:00 would all be matched with the meteorological data from 1/26/2014 16:00 to form a combined dataset. In addition to the GHI temporal or spatial standard deviations in the last 15-min interval, the GHI temporal or spatial standard deviations from the previous three 15-min intervals are also included in the predictor dataset. These four GHI standard

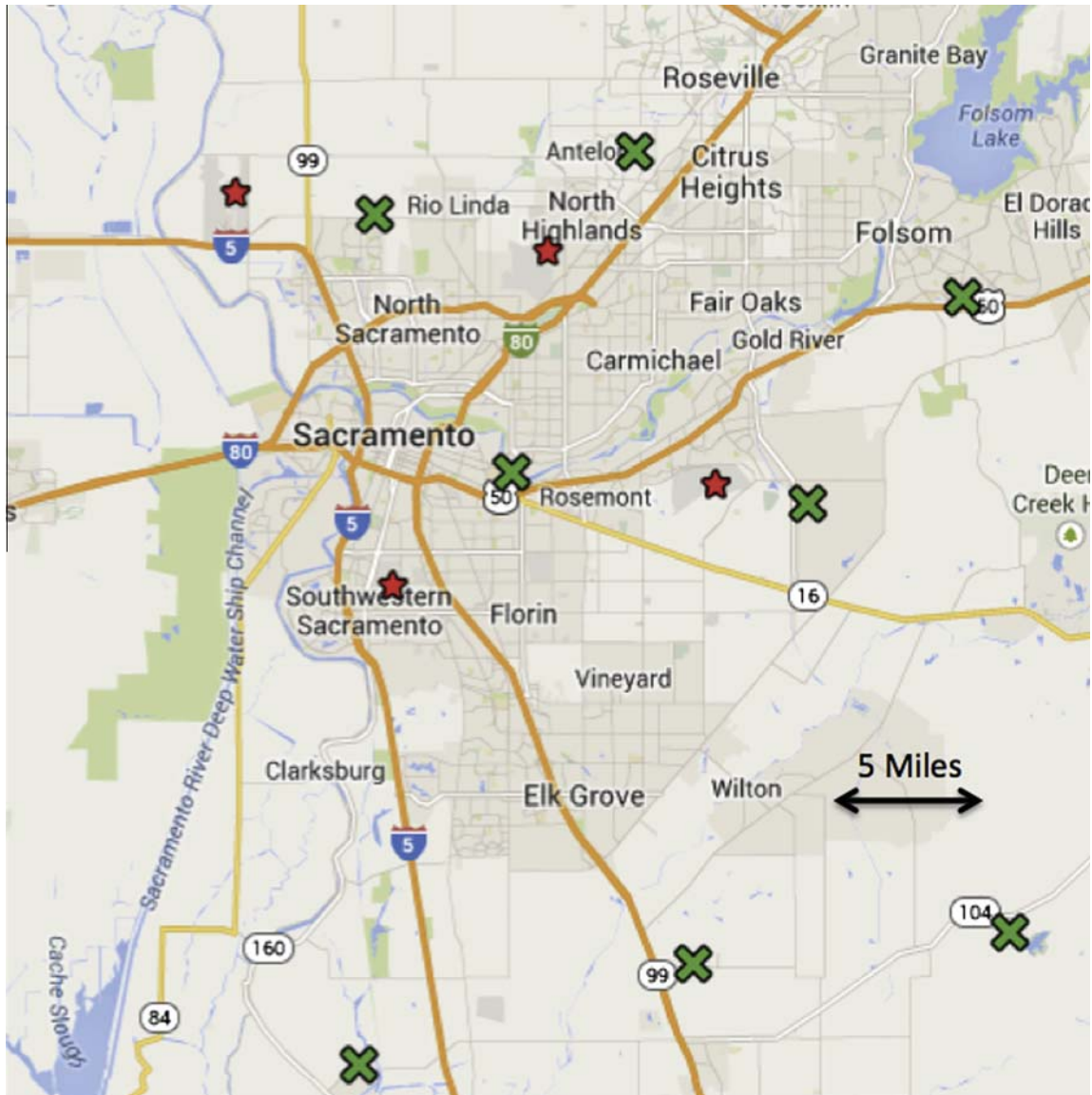


Fig. 1. SMUD observations locations shown as green X's with METAR observation locations shown as red stars. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

deviations for the 15-min intervals are provided as predictors so that the forecasting technique can model the recent trend in variability. Using the prior four 15-min intervals is appropriate because SMUD reports data on an hourly basis, so this study matches what will be implemented operationally. Fig. 2 shows a diagram of the predictors, which are the previous observations, observed weather data, and time information, that are fed into the model tree to predict the variability. The temporal standard deviation datasets include 40,127 instances combined for all eight SMUD sites. The spatial standard deviation dataset consists of 4057 instances aggregated at all eight sites. Instances where one or more location had missing data were omitted from the spatial standard deviation training and testing datasets.

Ten fold cross-validation randomly partitions the data into ten subsets to be used for training and testing the

model and provide an assessment of how the model tree will generalize to an independent set of data. The training of the model tree is performed on nine of ten subsets and the remaining subset is used as validation. This process is repeated for all of the ten subsets and the errors are averaged over the ten repetitions to reduce variability in the results.

3. Forecasting techniques

3.1. Baseline technique – climatology

The rationale for forecasting the temporal and spatial GHI standard deviation is to quantify the expected solar variability for utility companies and system operator's situational awareness of the expected irradiance variability over a 15-min timeframe. Thus, we seek to improve upon

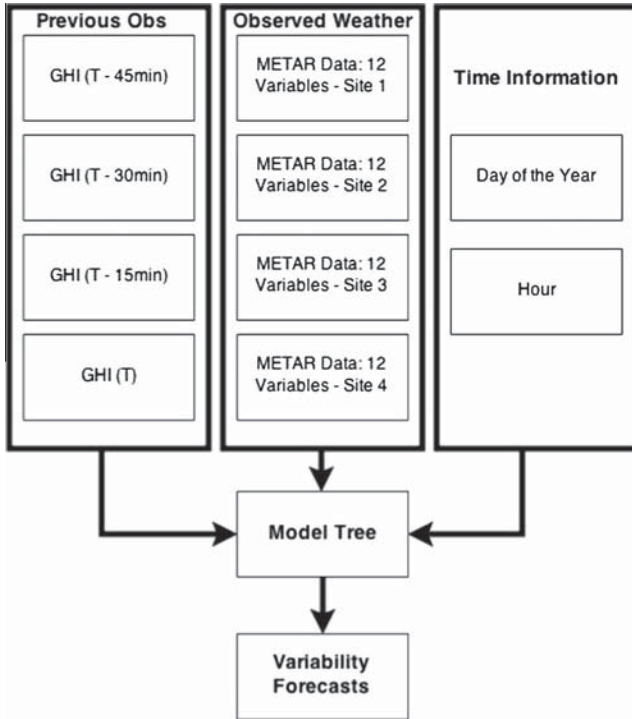


Fig. 2. Prediction flowchart that includes GHI and METAR Observations as well as time and date information.

the climatological mean values of the temporal and spatial GHI standard deviation at each forecast lead time. The spatial standard deviation mean value is calculated in a similar way by computing the training dataset's average for each 15-min forecast time interval out to 180 min.

$$STDEVMEAN_{\text{spatial_prediction}}(t) = \left(\sum_d^D \sqrt{\frac{\sum_{x_i}^x (GHI(x_i) - GHI_{s_mean})^2}{n}} \right) / D. \quad (3)$$

The n is equal to eight for the number of SMUD sites, the summation is over the number of training dataset instances (D). For example, to make a prediction for the 180-min spatial standard deviation, the climatological mean value of the 180-min spatial standard deviation in the training data is used as the prediction.

3.2. Model tree – cubist

The objective of a forecasting system is to model the actual relationships between the predictors and the predictand. In the case of weather forecasting, the relationship between the predictors and the predictand is frequently non-linear. Thus, a non-linear artificial intelligence prediction technique is often used. The artificial intelligence technique used in this study is the model tree, or Cubist model, which is Quinlan's (1992) M5 model tree formatted as a set of rules (Kuhn et al., 2012). The model tree uses a "separate-and-conquer" algorithm to search for a rule that explains part of the training instances, separates these

instances, and continues this process until no instances remain (Quinlan, 1996). The algorithm reformulates the tree into a set of rules and places a multivariate linear model at each leaf in order to predict our continuous predictands of solar irradiance variability. See Quinlan (1987a, 1987b, 1992) for a detailed explanation of this process. The process grows a tree that has multivariate linear regression models at its nodes and leaves. The final prediction is a weighted average of the multivariate linear regression equations at each node in the tree down to the final leaf (Kuhn et al., 2012). This weighted averaging is accomplished by a smoothing process that adjusts the predicted value from the leaf up to the root via,

$$PV(S) = \frac{n \times PV(S_i) + k \times M(S)}{n + k} \quad (4)$$

where n is the number of instances, i , in the node S , $PV(S_i)$ is the predicted value at node S and instance i S_i , k is a smoothing factor set equal to 15, and $M(S)$ is the model prediction at the leaf of the subtree. This smoothing is done to capture the skill in the predicted values at nodes along the tree down to the final leaf. Fig. 3 displays an example description of the model tree with the red branch highlighting the subtree used in this example prediction. Thus, this model tree is a set of rules that are paths from the top to bottom of the tree with each node's multivariate linear model output used in the final prediction.

We use an additional model option that combines the model tree's prediction with a prediction given the training dataset's nearest neighbor to further reduce the model tree's error. The nearest neighbor option first finds the training cases that are more similar to the current instance. Then the model tree is used to make predictions for all of the nearest neighbor instances and the current instance. The value of the current instance prediction is adjusted based on the difference between the current instance prediction and the prediction for the nearest neighbor instances. A sensitivity study, shown in Table 1, indicates that optimal configuration of the model tree includes one nearest neighbor, which results in the lowest mean absolute error for all four forecast lead times tested.

Separate model trees are configured for the temporal standard deviation and the spatial standard deviation separately. In addition, model trees are built for each lead-time for a total of 24 model tree configurations. The configurations of the model trees in this study have 100 rules. A sensitivity study (not shown) had minor and inconsistent differences in model error when the rules varied between 50 and 500.

4. Results

4.1. Temporal variability prediction results

To better understand the utility of our irradiance variability forecasts, we compute the Mean Absolute Error (MAE) for the prediction of the GHI temporal standard

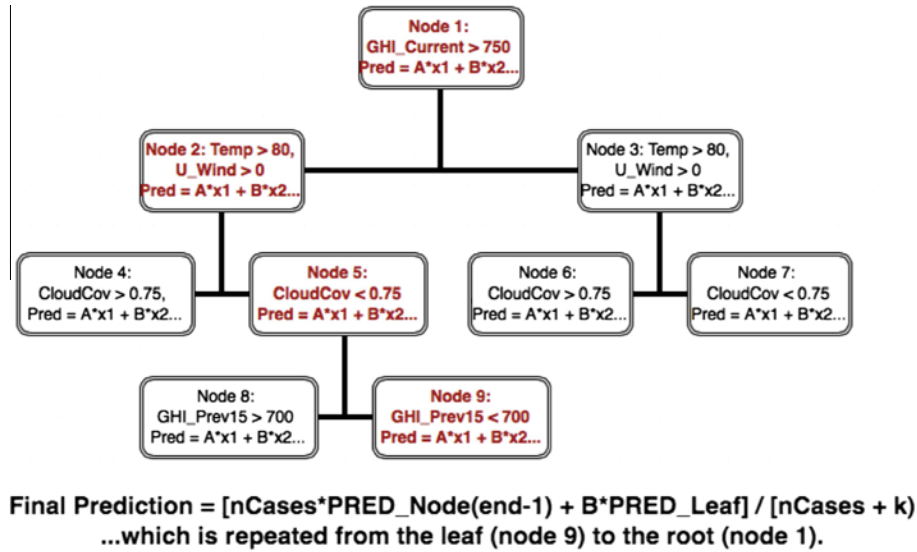


Fig. 3. Flowchart and description of the model tree. For this arbitrary instance, the subtree in red is used to make the final prediction via the equation at the bottom.

Table 1
Sensitivity study for the number of nearest neighbors used in the model tree prediction. The values shown are the MAEs of the model predicting GHI temporal standard deviation (W/m^2). Using one nearest neighbor results in the lowest MAE for all lead times.

All data – number of nearest neighbors – GHI					
Lead time	0	1	2	5	9
15	35.04	26.86	29.04	31.98	33.45
60	49.33	35.75	38.69	43.25	45.66
120	58.91	42.31	45.44	50.45	53.50
180	59.81	44.60	47.51	52.38	52.38

deviation and then compare this value to the error from assuming climatological averages. The MAE is computed as the average of the absolute differences between the forecast standard deviation of the GHI and the actual standard deviation of the GHI. Fig. 4 plots both the MAE of the spatial standard deviation of the GHI and the temporal standard deviation of the GHI for all forecast lead times. The model tree MAE for the GHI temporal standard deviation prediction increases from $16 W/m^2$ to $18 W/m^2$ as the forecast lead time increases from 15-min to 180-min. The results are similar for the spatial standard deviation prediction with errors ranging from approximately $15 W/m^2$ at 15-min lead-time to $20 W/m^2$ at 180-min lead-time.

In order to quantify the model tree forecast performance versus a baseline, the relative error is plotted in Fig. 5. The relative error is the error for the model tree divided by the error from climatology. Climatology is computed as the training dataset’s mean value of the GHI temporal standard deviation at that forecast lead time (t). Specifically, this is

$$Relative\ error(t) = \frac{Mean\ absolute\ error_{Cubist}(t)}{Mean\ absolute\ error_{mean_value}(t)} \quad (5)$$

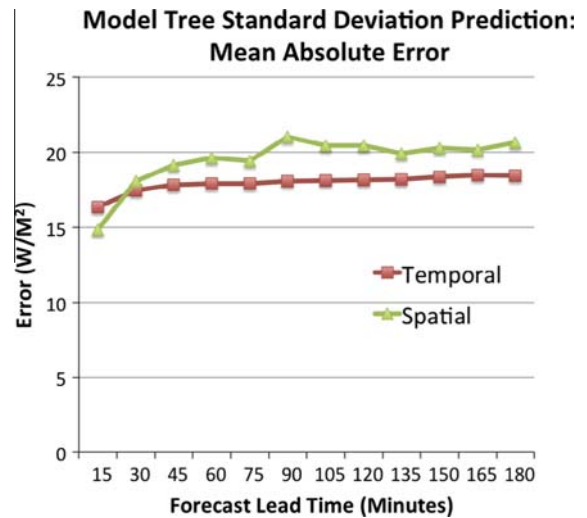


Fig. 4. Mean Absolute Error (MAE) for the model tree on the spatial standard deviation of the GHI (green – triangles) and temporal standard deviation of the GHI (red – squares). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The degree to which the relative error is less than 1.0 quantifies the forecasting skill improvement by the model tree compared to the climatological prediction. A value less than 1.0 indicates the model improves upon the baseline method of climatology. The relative error for the model tree begins at approximately 0.57 for the 15-min forecast lead time and increases slightly to a maximum value around 0.62 at 180-min forecast lead time. These results provide evidence that the model tree is approximately twice as accurate as using the climatological average value, thus providing utility companies with substantially more accurate forecasts of variability for resource management decision making.

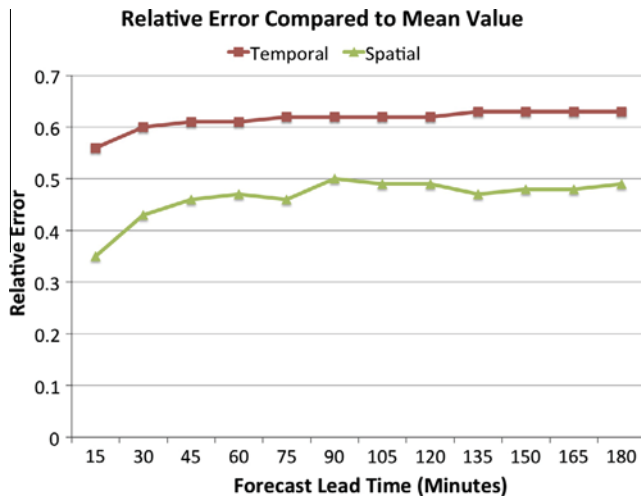


Fig. 5. Relative error for the model tree on the spatial standard deviation of the GHI (green – triangles) and temporal standard deviation of the GHI (red – squares). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Spatial variability prediction results

Next, we analyze the model tree's predictive ability for GHI spatial variability by examining the predictive skill of the standard deviation of the GHI among the SMUD observation sites. The MAE for the GHI spatial standard deviation prediction increases with forecast lead time as did that of the GHI temporal standard deviation prediction (Fig. 4). However, the error range over the forecast lead times is greater than that for the temporal data. Values of the MAE range from approximately 15 W/m^2 at 15 min to 21 W/m^2 at 180 min.

The relative error of the model tree compared to climatology (the mean GHI spatial standard deviation computed on the training dataset) is plotted in Fig. 5. The relative error for the model tree begins at approximately 0.35 for the 15-min forecast lead time and levels off at about 0.50 for forecast lead times longer than 75-min. This relative error provides evidence that the model tree is able to provide utility companies with at least twice the accuracy as assuming climatological average variability. This is a meaningful result for utility companies that have regional coverage with a range of distributed rooftop solar and solar power farms because the model tree is able to provide a substantial increase in the accuracy of predicting short range solar radiation variability across a region.

5. Evaluation of observed surface weather data as predictors

It is important to quantify the value of the surface weather observations as input into the model tree for utility companies to understand the value surface weather observation sites add to forecasting at solar power arrays. Therefore, an analysis was performed to compare the model tree without observed surface weather as predictors to including the observed surface weather variables as

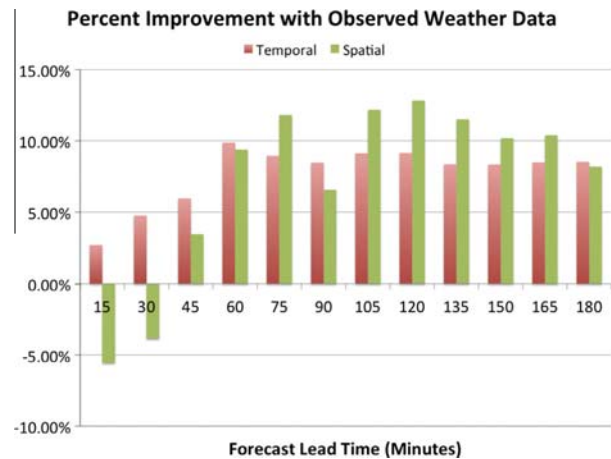


Fig. 6. Percent improvement with the model tree using the observed meteorological data as input compared to the cubist model without the observed meteorological data.

predictors. This is similar methodology to the data denial experiments of Kelly et al. (2007). The model tree trained without observed weather data is given six predictors: the last four 15-min GHI spatial or temporal standard deviation observations, the forecast hour, and the day of the year. The comparison between model tree results when trained with versus without observed weather as predictors is plotted in Fig. 6. The percent improvement is the percentage of error reduction when the model tree includes observed surface weather observations as predictors. With the exception of the first two lead times for the spatial standard deviation prediction, all other lead times show an improvement in skill when the observed weather data are included as predictors. The improvement steadily increases with lead time for the temporal standard deviation prediction until the 60-min lead time where the improvement levels off at around an 8% increase. After the first two forecast lead times, the improvement in GHI spatial standard deviation prediction with the observed weather used as predictors varies between 3% and 15% with an average percent improvement of about 10%. The negative percent improvement for the first two lead times may be due to the model tree over-fitting the most recent GHI spatial standard deviation values at the short range forecast lead times. The day of the year and hour have a higher contribution to the regression equations at longer forecast lead-times while the 15-min and 30-min forecast lead-times rely primarily on the last observed spatial GHI, which leads to large errors when the variability is highest. These results indicate that the model tree gains substantial predictive skill when observed weather is included as predictors.

6. Case study

We evaluate the model tree under weather conditions that are challenging to forecast. We examine the prediction on a day with morning clouds and afternoon sun for the temporal GHI standard deviation prediction (April 13th,

2014), and a mostly cloudy day across the SMUD irradiance observation sites for the spatial GHI standard deviation prediction (April 26th, 2014).

6.1. Temporal variability prediction case study

There are several forecasting challenges and successes shown when we examine a day of morning clouds followed by mostly clear conditions for predicting the GHI temporal variability at forecast lead-times of 15-min, 60-min, and 180-min. The predictions (blue lines) and the observed values (red lines) are plotted for the 15-min, 60-min, and 180-min forecast lead-times in Figs. 7–9. The x-axis is the forecast valid time in Universal Time Coordinate (UTC), which correspond to the daylight hours for April 13th in Sacramento, California.

At the 15-min lead-time (Fig. 7), the predictions follow closely to the actual observation; however, the predictions have a time lag. As shown previously, the observed surface weather from the METAR sites on the model tree improved prediction by only approximately ten percent. Therefore, one expects that the forecast would have a time lag because the model depends more on the most recent changes in GHI temporal standard deviations more than the observed meteorological data. When the irradiance variability changes quickly, the model needs the most recent information to update the forecast. The plots of the predictions for 60-min and 180-min lead-time exhibit a loss of skill compared to the 15-min lead-time. This loss of skill is similarly due to the error from the time lag between the variability measurement to the forecast valid time. However, when the irradiance variability is not rapidly changing, the model tree predictions at all lead

times are close to the clear day GHI values after the morning cloudiness dissipates.

6.2. Spatial variability prediction case study

We examine the ability of the model tree technique to predict the spatial variability of the GHI on a mostly cloudy day in the Sacramento area (April 26th, 2014). Similar to the temporal variability prediction plots, Figs. 10–12 are the spatial variability prediction plots for the 15-min, 60-min, and 180-min lead-times respectively.

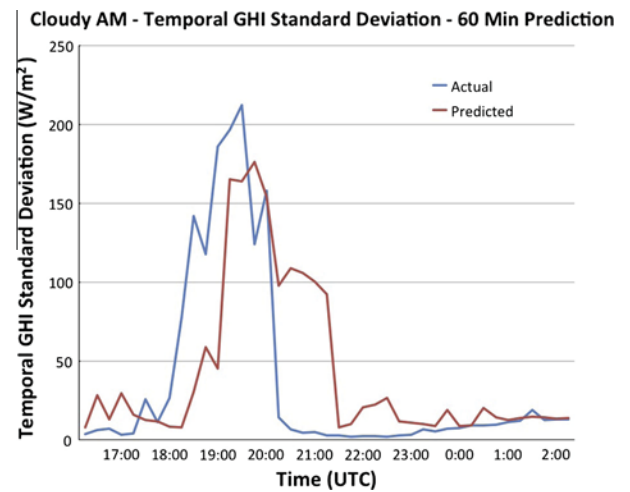


Fig. 8. Case study for the prediction of the temporal GHI standard deviation at a lead time of 60-min. The forecast valid time for the daylight hours of April 13, 2014 is plotted on the x-axis with the actual temporal GHI standard deviation values shown in blue and the predictions shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

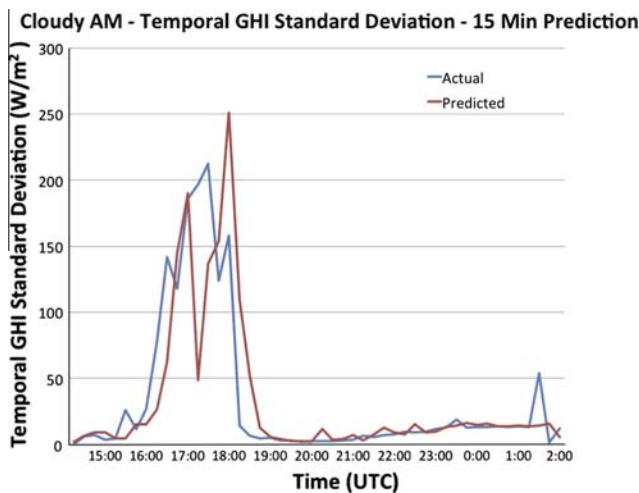


Fig. 7. Case study for the prediction of the temporal GHI standard deviation at a lead time of 15-min. The forecast valid time for the daylight hours of April 13, 2014 is plotted on the x-axis with the actual temporal GHI standard deviation values shown in blue and the predictions shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

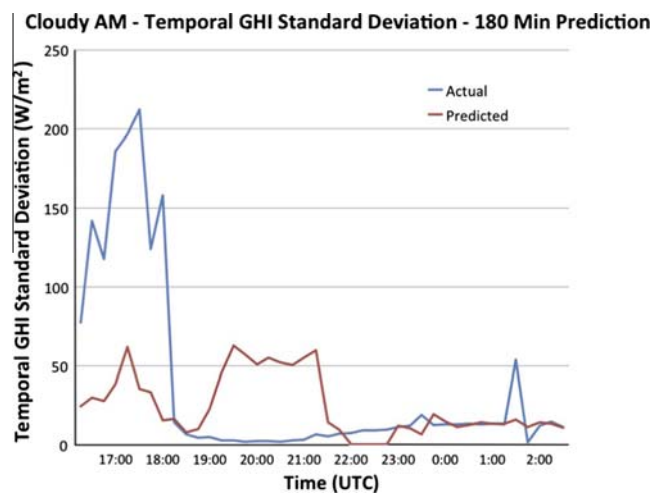


Fig. 9. Case study for the prediction of the temporal GHI standard deviation at a lead time of 180-min. The forecast valid time for the daylight hours of April 13, 2014 is plotted on the x-axis with the observed temporal GHI standard deviation values shown in blue and the predictions shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

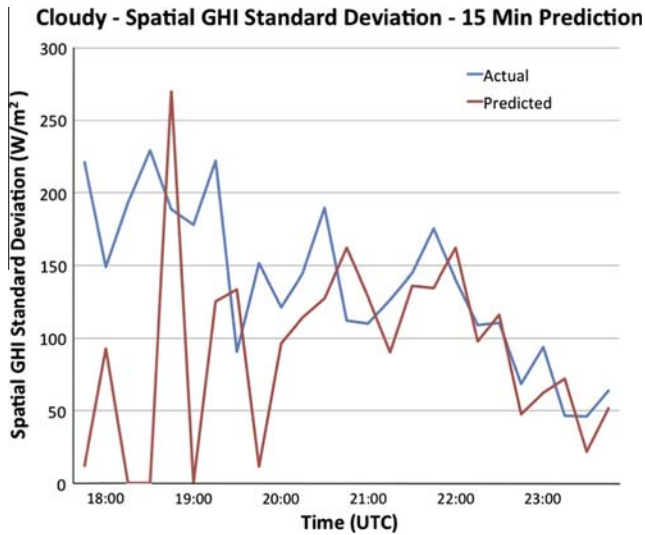


Fig. 10. Case study for the prediction of the spatial GHI standard deviation at a lead time of 15-min. The forecast valid time for the daylight hours of April 13, 2014 is plotted on the x-axis with the observed spatial GHI standard deviation values shown in blue and the predictions shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

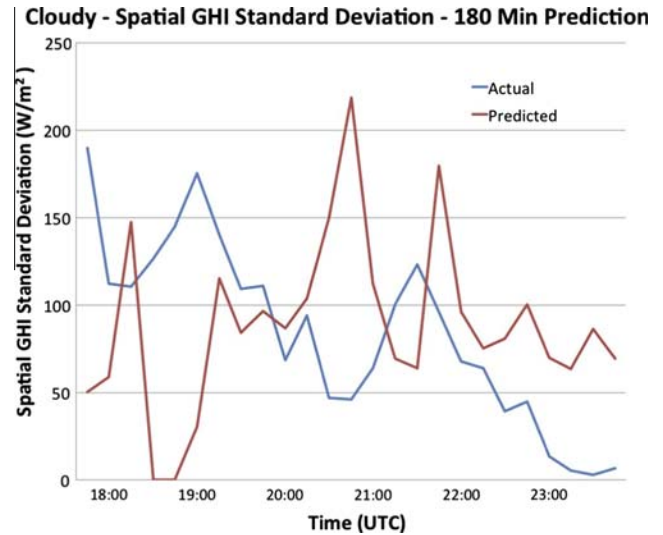


Fig. 12. Case study for the prediction of the spatial GHI standard deviation at a lead time of 180-min. The forecast valid time for the daylight hours of April 13, 2014 is plotted on the x-axis with the observed spatial GHI standard deviation values shown in blue and the predictions shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

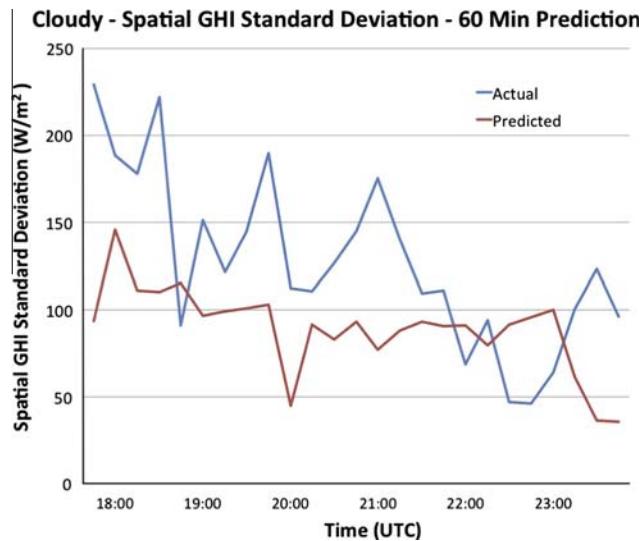


Fig. 11. Case study for the prediction of the spatial GHI standard deviation at a lead time of 60-min. The forecast valid time for the daylight hours of April 13, 2014 is plotted on the x-axis with the observed spatial GHI standard deviation values shown in blue and the predictions shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The 15-min spatial GHI standard deviation forecasts experience a similar time lag to the 15-min temporal GHI standard deviation forecasts. This again highlights the greater relative importance of the recent spatial GHI standard deviation observations compared to the surface weather observations as input to the model tree. The 60-min spatial GHI standard deviation forecasts show an average under-prediction of the spatial variability over the entire

day. This is potentially due to having too few overcast days to have a reasonable sample for the model tree to accurately predict the overcast forecast lead times. The 180-min spatial GHI standard deviation forecasts are only able to forecast the overall trend of the GHI spatial variability in the last two hours of the forecast period. Once again, this is likely due to lacking a reasonable sample of similar training cases with cloudy conditions across the Sacramento area.

7. Discussion and conclusions

A significant challenge with utilization of solar energy is its variable nature; therefore, the focus of this study is evaluating whether we can accurately predict the temporal and spatial solar irradiance variability for the Sacramento area. The variability was quantified via the 15-min temporal standard deviation of the GHI and the spatial standard deviation of the GHI across irradiance observation sites. The model tree artificial intelligence algorithm with a nearest neighbor option was trained on data both from the METAR weather observations and from Sacramento area irradiance measurements. Short range predictions were made at 15-min intervals out to 180 min for both the temporal and spatial standard deviations of the GHI.

The predictive ability of the model tree was assessed using the MAE of the model and the relative error compared to using the mean, over the training dataset, of the temporal or spatial standard deviation of GHI (i.e. variability climatology). This mean value is computed separately for each forecast lead time. The relative error showed that for both spatial and temporal variability, the model tree technique is able to produce forecasts with

approximately half the error of the climatological variability forecast. The case studies of the spatial and temporal variability highlighted the importance of the most recent GHI variability observations, and the need for GHI variability observations at multiple times in order to capture trends.

Because solar energy is inherently a highly variable renewable energy resource while stability of the energy distribution network is essential, the added value of accurate GHI variability prediction is significant for utility companies and system operators. As solar energy penetration continues to grow in many markets across the United States, these entities will require the estimation of near term solar resource variability. The results for the model tree technique indicate that could be a beneficial technique for utility companies to implement in real-time forecasting of short range solar irradiance variability.

Future work will test other, probabilistic, approaches to forecasting of the GHI and its variability. A longer dataset and data from additional regions in the United States will provide a thorough evaluation of the GHI temporal and spatial variance prediction with the model tree.

Disclaimer

“This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.”

Acknowledgements

This material is based upon work supported by the Department of Energy under SunCast Award Number [DE-EE0006016]. We gratefully acknowledge all of the collaborators on the SunCast project for insightful discussions and ideas, including Seth Linden, Sheldon Drobot, Jared Lee, Julia Pearson, Laura Hinkelman, and Tara Jensen. This project would not have been possible without the data from the Sacramento Municipal Utility District and the help from Thomas Brummet at NCAR for the data quality control and processing. The authors also thank these anonymous reviewers whose thoughtful comments have resulted in an improved paper.

References

- Almonacid, F., Pérez-Higueras, P.J., Fernández, E., Hontoria, L., 2014. A methodology based on dynamic artificial neural network for short-term forecasting of the power output of a PV generator. *Energy Convers. Manage.* 85, 389–398.
- Bartholomy, O., Vargas, T., Simone, M., Hansen, C., Fitchett, S., Pohl, A., 2014. Benchmarking solar power and irradiance forecasting accuracy at Sacramento municipal utility district. *IEEE Photovoltaic Specialist Conference*, June 8–13, pp. 63–68.
- Bhardwaj, S., Sharma, V., Srivastava, S., Sastry, O.S., Bandyopadhyay, B., Chandel, S.S., Gupta, J.R.P., 2013. Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model. *Sol. Energy* 93, 43–54.
- Bouzerdoum, M., Mellit, A., Pavan, A., 2013. A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Sol. Energy* 98, 226–235.
- Chu, Y., Pedro, H., Coimbra, C., 2013. Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning. *Sol. Energy* 98, 592–603.
- Curtright, A.E., Apt, K., 2008. The character of power output from utility-scale photovoltaic systems. *Prog. Photovoltaics Res. Appl.* 16, 241–247.
- Diagne, M., David, M., Lauret, P., Boland, J., Schmutz, N., 2013. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renew. Sustain. Energy Rev.* 27 (11), 65–76.
- Ela, E., Diakov, V., Ibanex, E., and M. Heaney, 2013: Impacts of Variability and Uncertainty in Solar Photovoltaic Generation at Multiple Timescales. Technical Report: NREL/TP-5500-58274.
- Fernandez, E., Almonacid, F., Sarmah, N., Rodrigo, P., Mallick, T.K., Perez-Higueras, P., 2014. A model based on artificial neuronal network for the prediction of the maximum power of a low concentration photovoltaic module for building integration. *Sol. Energy* 100, 148–158.
- Fu, C.-L., Cheng, H.-Y., 2013. Predicting solar irradiance with all-sky image features via regression. *Sol. Energy* 97, 537–550.
- Gueymard, C.A., Wilcox, S.M., 2011. Assessment of spatial and temporal variability in the US solar resource from radiometric measurements and predictions from models using ground-based or satellite data. *Sol. Energy* 85, 1068–1084.
- Hinkelman, L.M., Evans, K.F., Clothiaux, E.E., Ackerman, T.P., Stackhouse Jr., P.W., 2007. The effect of cumulus cloud field anisotropy on domain-averaged solar fluxes and atmospheric heating rates. *J. Atmos. Sci.*, 3499–3520
- Hinkelman, L.M., 2013. Relating Solar Resource Variability to Cloud Type. NREL Progress Report.
- Hoff, T.E., Perez, R., 2012. Modeling PV fleet output variability. *Sol. Energy* 86, 2177–2189.
- Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. Solar forecast methods for renewable energy integration. *Prog. Energy Combust. Sci.* 39, 535–576.
- Kelly, G., Thepaut, J.-N., Buizza, R., Cardinali, C., 2007. The value of observations. I: Data denial experiments for the Atlantic and the Pacific. *Q.J.R. Meteorol. Soc.* 133, 1803–1815.
- Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2012. Cubist Models for Regression. <<http://rulequest.com/cubist-info.html>>.
- Kuzmaul, S., Ellis, A., Ellis, A., Stein, J., Johnson, L., 2010. Lanai high-density irradiance sensor network for characterizing solar resource variability of MW-scale PV system. In: 35th IEEE PVSC, Honolulu HI.
- Lave, Matthew, Kleissl, Jan, Stein, Joshua S., 2013. A wavelet-based variability model (WVM) for solar PV power plants. *IEEE Trans. Sustain. Energy* 4 (2), 501–509.
- Lew, D., Brinkman, G., Ibanez, E., Hummon, M., Hodge, B.-M., Heaney, M., 2012. Sub-hourly impacts of high solar penetrations in the western United States. In: 2nd Annual International Workshop on Integration of Solar Power into Power Systems Conference, Lisbon, Portugal Nov 12–13, 2012.

- Mahoney, W.P., Parks, K., Wiener, G., Yubao, L., Myers, W.L., Juanzhen, S., Delle Monache, L., Hopson, T., Johnson, D., Haupt, S.E., 2012. A wind power forecasting system to optimize grid integration. *IEEE Trans. Sustain. Energy* 3 (4), 670–682.
- Marquez, R., Gueorguiev, V.G., Coimbra, C.F.M., 2011. Forecasting of global horizontal irradiance using sky cover indices. In: Proceedings of the ASME 2011 5th International Conference on Energy Sustainability. Washington DC, Aug. 7–10 2011.
- Martin, L., Zarzalejo, L.F., Polo, J., Navarro, A., Marchante, R., Cony, M., 2010. Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning. *Sol. Energy* 84, 1772–1781.
- McCandless, T.C., Haupt, S.E., Young, G.S., Annunzio, A.J., 2015. A regime dependent bayesian approach to short-term solar irradiance forecasting. In: 13th Conference on Artificial Intelligence: Applications of Artificial Intelligence Methods for Energy-Part II. Phoenix, AZ, Amer. Meteor. Soc. J6.5.
- Mellit, A., 2008. Artificial intelligence technique for modeling and forecasting of solar radiation data: a review. *Int. J. Artif. Intell. Soft Comput.* 1 (1), 52–76.
- Mills, A., Ahlstrom, M., Brower, M., Ellis, A., George, R., Hoff, T., Kroposki, B., Lenox, C., Miller, N., Stein, J., Wan, Y., 2009. Understanding variability and uncertainty of photovoltaics for integration with the electric power system. *Electr. J. (LBNL-2855E)*.
- Nguyen, D., Kleissl, J., 2014. Stereographic methods for cloud base height determination using two sky imagers. *Sol. Energy* 107, 495–509.
- Pedro, H.T.C., Coimbra, C.F.M., 2012. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Sol. Energy* 86, 2017–2028.
- Perez, R., David, M., Hoff, T., Kivalov, S., Kleissl, J., Lauret, P., Perez, M., 2015. Spatial and temporal variability of solar energy. *Found. Trends Renew. Energy* (in press).
- Quesada-Ruiz, S., Linares-Rodríguez, A., Ruiz-Arias, J.A., Pozo-Vázquez, D., Tovar-Pescador, J., 2015. An advanced ANN-based method to estimate hourly solar radiation from multi-spectral MSG imagery. *Sol. Energy* 115, 494–504.
- Quinlan, J.R., 1987a. Generating production rules from decision trees. In: Proceedings of the Tenth International Joint Conference of Artificial Intelligence, Milan, Italy, pp. 304–307.
- Quinlan, J.R., 1987b. Simplifying decision trees. *Int. J. Man Mach. Stud.* 27, 221–234.
- Quinlan, J.R., 1992. *Learning with Continuous Classes*. World Scientific.
- Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. *J. Artif. Intell. Res.* 4, 77–90.
- Reikard, G., 2009. Predicting solar radiation at high resolutions: a comparison of time series forecasts. *Sol. Energy* 83, 342–349.
- Rayl, J., Young, G.S., Brownson, J.R.S., 2013. Irradiance co-spectrum analysis: tools for decision support and technological planning. *Sol. Energy* 95, 354–275.
- Witten, I.H., Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
- Zagouras, A., Inman, R., Coimbra, C.F.M., 2014. On the determination of coherent solar microclimates for utility planning and operations. *Sol. Energy* 102, 173–188.