A regime-dependent artificial neural network technique for short-range solar irradiance forecasting

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, USA
[b] The Pennsylvania State University, Department of Meteorology, 503 Walker Building, University Park, PA 16802-5013, USA

ABSTRACT

Solar power can provide substantial power supply to the grid; however, it is also a highly variable energy source due to changes in weather conditions, i.e. clouds, that can cause rapid changes in solar power output. Independent systems operators (ISOs) and regional transmission organizations (RTOs) monitor the demand load and direct power generation from utilities, define operating limits and create contingency plans to balance the load with the available power generation resources. ISOs, RTOs, and utilities will require solar irradiance forecasts to effectively and efficiently balance the energy grid as the penetration of solar power increases. This study presents a cloud regime-dependent short-range solar irradiance forecasting system to provide 15-min average clearness index forecasts for 15-min, 60-min, 120-min and 180-min lead-times. A k-means algorithm identifies the cloud regime based on surface weather observations and irradiance observations. Then, Artificial Neural Networks (ANNs) are trained to predict the clearness index. This regime-dependent system makes a more accurate deterministic forecast than a global ANN or clearness index persistence and produces more accurate predictions of expected irradiance variability than assuming climatological average variability.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

The proliferation of photovoltaic (PV) power production has made accurate short-range solar irradiance forecasts a necessity for utility companies to ensure reliable and efficient integration of solar power into the energy grid. An accurate forecast for irradiance is necessary; however, a prediction of the variability of the irradiance is also helpful in maintaining reliable energy production with increased levels of solar power integration. The amount of solar irradiance reaching the PV panels depends on both the diurnal cycle and on the atmospheric state. While the diurnal cycle is easily forecast, the stochastic element of cloud formation makes that component of irradiance forecasting a challenge. This problem can be mitigated to some extent if one can forecast the cloud type. The identification of cloud types, i.e. cloud regimes, is a valuable tool in short-range solar irradiance forecasting because each cloud regime is associated with particular cloud properties such as cloud optical

depth, cloud growth rate, and cloud dissipation rate; and therefore, have various degrees of irradiance attenuation. The cloud type also impacts the short-range temporal and spatial irradiance variability and the corresponding irradiance forecast uncertainty.

The optimal method for solar irradiance prediction depends on several factors, including the forecast lead time, with statistical techniques and cloud advection techniques most effective for short-range irradiance forecasting. Short-range forecasting is defined here as solar irradiance predictions from 15 min out to 180 min. Predicting solar power through statistical techniques has gained the attention of researchers in recent years [22]. found that a Support Vector Machine approach to post-processing Numerical Weather Prediction (NWP) Models' forecasts produced lower Global Horizontal Irradiance (GHI) forecast error compared to linear regression post-processing techniques [8] and [25] found that AutoRegressive Integrated Moving Average (ARIMA) models produced lower solar irradiance and solar power errors compared to other time-series short-range prediction techniques while [17] used a Markov process to predict sunshine and cloud cover [15]. reported that Artificial Neural Networks (ANNs) have been used in modeling and predicting solar radiation more than any other non-linear technique. More recently, several studies determined that

* Corresponding author. National Center for Atmospheric Research, 3450 Mitchell Lane, Boulder, CO 80301, USA.
E-mail addresses: mccandle@ucar.edu (T.C. McCandless), haupt@ucar.edu (S.E. Haupt), g3y@psu.edu (G.S. Young).

models based on ANNs improve solar irradiance or solar power forecast accuracy compared to various baseline techniques [2,3,6,12,14,23]. Several studies have examined the performance of these statistical forecast models in various weather conditions [18]. found the accuracy of an ANN optimized with a Genetic Algorithm had a strong seasonal dependence [13]. correlated total sky images, infrared data, and solar radiation observations at the surface to use as input to an ANN and found the variability of solar radiation to be strongly dependent on the amount of cloud cover. Each day was classified as sunny, partly sunny or cloudy and an ANN was used to forecast the daily profile of the power produced by a PV plant [16]. [4] concluded that the ANN model has lower errors for days characterized by direct irradiance (clear days) and for days characterized by diffuse irradiance (cloudy days) than for days characterized by a mix of direct and diffuse irradiance (partly cloudy days).

This current work seeks to improve two major facets of short-range solar irradiance forecasting via regime-dependent statistical forecasting: deterministic irradiance forecast accuracy and irradiance variability estimates. We first classify cloud regimes with a k-means algorithm and then apply ANNs to each regime to produce a more accurate GHI forecast with variability estimates. The GHI prediction is generally a necessary step in the prediction process before converting to power. The k-means algorithm statistically classifies the cloud regime based on surface weather and irradiance observations. This approach parallels that of [5]; who classified weather regimes with Principal Component Analysis (PCA) in order to apply regime-dependent optimal weights to ensemble temperature forecasts. After k-means clustering, ANNs are implemented for each weather regime independently with the intention of modeling each weather regime's inherent predictability, and thus, each regime's different causal relationships between predictors and predictand. Predictions are made for the clearness index (Kt), which is the ratio of the observed GHI at the surface to the Top Of Atmosphere (TOA) expected GHI. The prediction of Kt is important for utility companies because it quantifies the amount of attenuation from aerosols and clouds at a particular location [13]. Although it is common to classify days at being cloudy, partly cloudy or clear, we classify cloud regimes that are specific to the short-range development of clouds and therefore improve the forecast error for lead-times up to 180-min.

We wish to make short-range predictions for multiple sites near Sacramento, California for 15-min intervals out to 180 min. In operational forecasting, these short-range predictions are blended with forecasts from NWP models and a satellite based cloud advection technique then converted to power in the National Center for Atmospheric Research SunCast System that predicts solar power out to 168-h [7].

Section 2 provides an overview of our approach. In section 3, we discuss the data and the additional predictors derived from the initial datasets, which are the Sacramento Municipal Utility District (SMUD) irradiance network and the METAR network. In section 4, the prediction techniques of the ANNs and the baseline clearness index persistence forecast are described. In section 5, we summarize the k-means algorithm for cloud regime classification and the selection of optimal inputs. We describe the prediction methods before the regime classification method because we use the ANN prediction method to inform our decision on the best selection of inputs for the k-means regime classification. In section 6, we present and discuss the prediction results. The final section, 7, summarizes and poses potential future work.

2. Approach

The goal of this work is to develop a cloud regime-dependent short-range solar irradiance forecast system in order to not only improve the deterministic forecast accuracy, but also to provide a quantification of the expected solar irradiance variability and corresponding forecast uncertainty. This section outlines our classify-then-predict process; the details are described in the following sections. Our methodology begins by classifying the cloud regime with the k-means algorithm. We then train a separate ANN to make predictions for each individual regime as depicted in Fig. 1. This novel work goes beyond [16] and [12]; and others in the sophistication and automation in identifying regimes with the k-means algorithm and in the regime-dependent configuration of the ANNs that are specific to improving the final prediction algorithm. The process begins by selecting the optimal set of inputs for cloud regime classification that corresponds to the final model with the lowest forecast error. The selected set of inputs is then used by the k-means algorithm to classify and partition the datasets into an optimal number of cloud regime subsets. Finally, ANNs are constructed on each of the cloud regime datasets independently. This classify-then-predict process (with k-means then ANN) is repeated for each forecast lead-time.

A cloud regime-dependent Artificial Intelligence (AI) system requires dividing the cases into distinct regimes for which the fundamental relationship between predictors and predictand is expected to differ, and therefore, to allow more accurate short-range forecasts. Thus, careful sensitivity studies determined the optimal configurations of the AI models in order to match the complexity of the relationships among the predictors in the regimes. After all data are quality controlled and additional variables are derived, the datasets are randomly split 2/3 for training and 1/3 for testing. All of the results shown are from the testing datasets; however, the sensitivity tests conducted to determine the optimal configurations of the system were performed on the training datasets. The ANN and k-means sensitivity studies similarly split the training dataset into 2/3 for training and 1/3 for testing and the optimal configuration was determined based on this 1/3 independent test set. This approach avoids compromising the independence of the initial test dataset. We show results in this study for four forecast lead-times: 15 min, 60 min, 120 min, and 180 min. These predictions are for the 15-min average clearness index
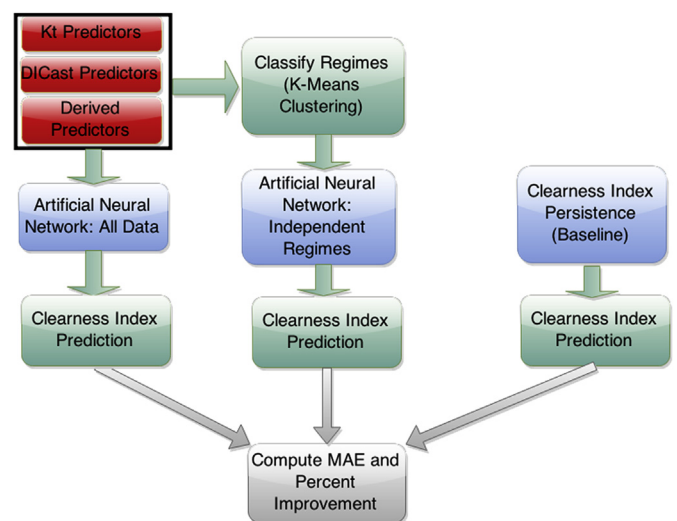

Fig. 1. Process design: first classify cloud regimes on the optimal set of potential inputs shown in the red rectangles outlines in the black box, then apply ANN models to predict the clearness index on each regime independently. An ANN is also applied on all data (i.e. without regime identification), and compared to the clearness index persistence prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ending at each lead-time.

In a real-time forecasting environment, predictions are made in a three-step process. In the preprocessing step, the data are collected, quality controlled, and the additional predictors are computed. In the regime classification step, the trained k-means algorithm is applied to the current data. Then, the ANN trained for the currently classified cloud regime and forecast lead-time is used to predict the solar irradiance (clearness index). The ANNs use numerical weather prediction analysis data and irradiance observations as input to predict clearness index at multiple locations in the vicinity of Sacramento, California.

3. Data

3.1. Irradiance data

The network of irradiance observing sites used in this study is that of the Sacramento Municipal Utility District (SMUD) in California, which is an area of varying terrain approximately 900 square miles [1]. We use data from eight solar power forecast sites that measure irradiance, shown in Fig. 2 as blue triangles. The GHI observations are available for a period of 367 days from January 25th, 2014 through January 26th, 2015. The temporal resolution of the raw data is one minute and averages are computed over 15-min intervals ending at :00, :15, :30, and :45 for each hour. The 15-min averaged GHI data are then converted to clearness index values. The clearness index is the ratio of the observed GHI at the surface to the Top Of Atmosphere (TOA) expected GHI, which is computed via a series of geometric calculations for a given location and time. This averaging interval was selected after communication with several utility companies and agrees with the shortest time range for which a forecast is currently useful for dispatch decision-making. The solar irradiance data from all eight solar power sites is aggregated and all instances with missing data or nighttime observations are excluded from the final dataset. There are a total of 71,184 instances in the final dataset.

In order to evaluate the prediction techniques, this study follows the same procedure as the planned real-time operational implementation. The data are provided from the utility company every hour with the one minute raw data averaged over 15-min intervals ending at :00, :15, :30, and :45 for each hour. Therefore, our prediction techniques ingest four predictors from the irradiance data, each converted to Kt: the average Kt from 60 to 45 min, from 45 to 30 min, from 30 to 15 min, and the past 15 min prior to forecast initialization time, which is the start of every hour. Hereafter, these predictors are named Kt_Prev60, Kt_Prev45, Kt_Prev30, and Kt_Prev15.

3.2. Weather data

The meteorological dataset used here is from the Meteorological Aviation Report (METAR) network, which represents hourly surface weather observations from stations typically located at airports across the United States. The METAR observations are quality controlled by the National Center for Atmospheric Research (NCAR) for ingest to the Dynamic Integrated foreCast (DICast) System [10]. The closest METAR sites to the SMUD irradiance observations sites are the three locations plotted as red X's in Fig. 2. We use six weather variables: cloud cover, dewpoint temperature, categorical precipitation in the last hour (1 = precipitation occurred, 0 = precipitation did not occur), precipitation amount, temperature, and wind speed.

3.3. Additional derived variables

In training AI methods, it is often useful to employ derived variables that emphasize important physical processes. Here, we derive inputs specific to the k-means classification system, as well
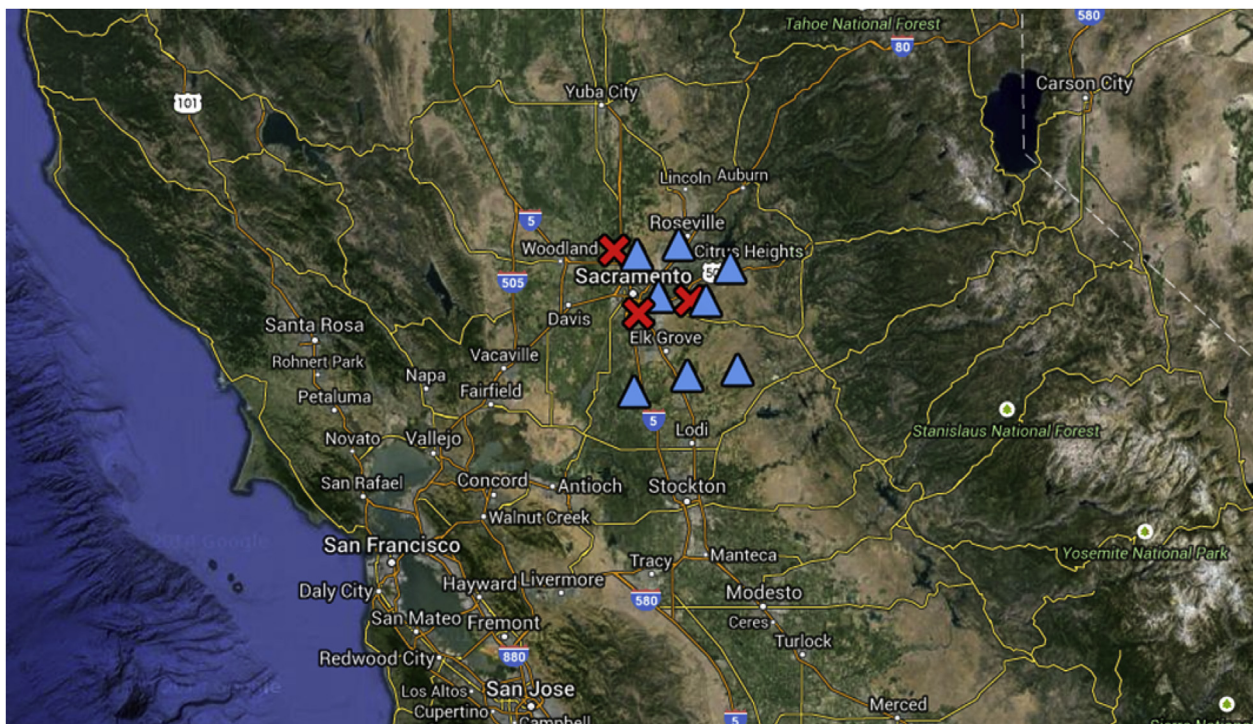

Fig. 2. Map of the SMUD sites (blue triangles) and METAR/DICast predictor sites (red X's). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predictors specific to the ANN prediction system. In particular, we leverage our meteorological knowledge to provide the k-means algorithm with the inputs that are specific to the goal of identifying cloud regimes and to provide the ANNs with predictors that are specific to the goal of solar irradiance prediction, i.e. the predictors most important for forecasting the evolution of clouds.

We first assess whether averaging the predictor values from all three METAR sites improves the forecast accuracy of the ANNs. A sensitivity study (not shown) revealed that providing the forecasting technique with data from each of the three sites produced higher overall skill of the prediction; thus, the data were not averaged. This result highlights the ability of the ANN to capture atmospheric relationships among predictors at the different locations that provide insight into the spatio-temporal nature of the evolving atmospheric state. Dewpoint depression, defined as the difference between the temperature and the dewpoint temperature at a height two meters above the ground, quantifies the atmosphere's nearness to saturation. This derived predictor is averaged over the three METAR sites after a sensitivity study showed no improvement by including the dewpoint depression for each site. Preliminary testing indicated that the cloud cover predictors have the highest importance in the GHI prediction, unsurprisingly [11]. Therefore, two predictors are derived from the cloud cover at the three sites. First, the variability is quantified by taking the standard deviation of the cloud cover across the three sites. Second, the mean of the cloud cover averaged at those sites is squared to emphasize the importance of thick, unbroken cloudiness in the region.

Predictors were also derived from the irradiance observed at the SMUD sites and the corresponding clearness index values. For instance, we computed the standard deviation of the four 15-min averages at the forecast site to quantify the irradiance variability rather than providing all four observations. Another predictor was derived by fitting a linear equation to these four data points and using its slope to capture the trend of the clearness index. Additionally, the most recent trend was quantified by computing the most recent change in clearness index by subtracting the Kt_Prev30 from the Kt_Prev15. The final derived predictors from the irradiance observations characterize the spatial distribution of the clearness index over the Sacramento Valley in the past 15 min. The mean and standard deviation of the nearby SMUD sites, i.e. the observations at the remaining seven sites not including the site being forecast for, were computed and used as predictors.

The last step in deriving predictors for the ANN was to transform time variables. The time was converted from UTC to local time to move the 24 to 0 discontinuity into the nighttime hours. The Julian Day was converted to the sine of the Julian Day to normalize the value corresponding to the day of the year between −1 and 1. Thus, the ANN was provided 32 predictors, including the NWP analysis predictors provided for all three DICast sites (column three), as shown in Table 1. Some of these derived predictors for the ANN model were also tested as derived inputs for the k-means regime

classification. The optimal selection of those inputs, including derived inputs, is described in detail in section 5.

4. Prediction techniques

4.1. Artificial neural network

We use the ANN as the non-linear AI prediction technique for our forecasts. ANNs' advantages include their ability to model non-linear processes without the assumption of the form of the relationship between input and output variables. In the review by Ref. [15]; AI models have been successfully developed to forecast solar radiation, clearness index, and insolation. [21] found that AI approaches significantly outperform traditional linear models in uni- and multi-variate studies, with the ANN feed-forward approach showing the best results.

The ANN used here is a feed-forward neural network trained by a backpropagation algorithm that is also known as a multi-layer perceptron [19,20]. The ANN, which is an algorithm that functions analogous to the human nervous system, is constructed of interconnected signal processing units (i.e. neurons) that calculate output values based on inputs and a set of weights and biases (i.e. multiplicative and additive scaling factors) which are tuned during the training process. The feed-forward neural network permits only forward connections. Fig. 3 is an example neural network where all the predictors are connected to each neuron and then each neuron is connected to the output layer that computes the final prediction. In our configuration, there are more predictors and typically more neurons than shown in Fig. 3; however, this diagram serves to show the ANN's information flow, from left (predictors) to right (output). The specific neural network module used in this study is the newff model in the Neurolab python library (https://pythonhosted.org/neurolab/). This ANN configuration has several tunable parameters and the optimal configuration was determined from multiple
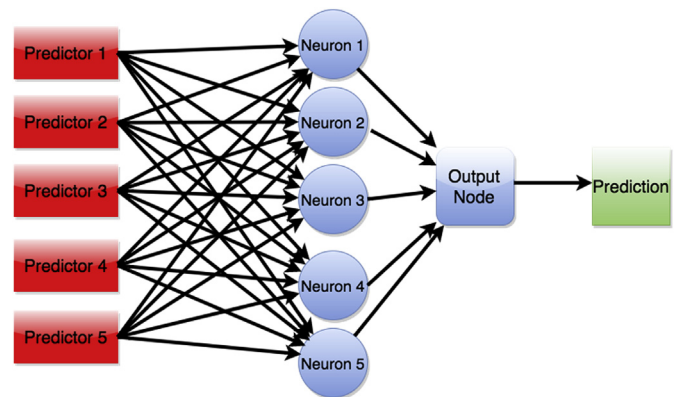

Fig. 3. Schematic of a feed-forward Artificial Neural Network used in this study.

Table 1
List of all the observed and derived predictors for the ANN. The Observed Weather column has three observations for each variable listed because there are three observation sites; thus, there are 32 total predictors.

Predictors				
Observed Kt	Derived Kt	Observed weather (3 sites)	Derived weather (averages)	Time
Kt_{Prev60}	Kt slope (1-hr)	Temperature	Dewpoint depression	Sin of the Julian Day
Kt_{Prev45}	Kt temporal varaibility (Stdev)	Dewpoint temperature	Cloud cover variability (Stdev)	Local hour
Kt_{Prev30}	Kt_{Prev15}- Kt_{Prev30}	Wind speed	Cloud cover squared	
Kt_{Prev15}	Kt Nearby mean	Cloud cover		
	Kt Nearby variability (Stdev)	Categorical precipitation (1-hr)		
		Precipitation amount (1-hr)		

sensitivity studies on subsets of the training data [24]. The optimal ANN configuration was determined to have one hidden layer, a learning rate of 0.01, and a weight decay of 0.5. The ANNs are trained for 200 epochs in order to adjust the weights and biases that minimize the error between the ANN outputs and the predictands without over-fitting the data. However, the number of neurons in the hidden layer is allowed to vary with regime because the regimes differ in the number of cases and in different levels of complexity in relationships between the predictors and the predictand. Therefore, the ANNs are trained with 5, 10, 15, 20, 25, or 30 hidden layer neurons and the configuration with the lowest error on a subset of the training data held out for an independent verification was the configuration chosen for that regime. The ANN is also used on the dataset without regime separation to provide the basis from which to quantify the improvement in forecast skill resulting from regime identification.

4.2. Clearness index persistence

We wish to compare all of our ANN-based prediction techniques to a baseline prediction method, for which we use clearness index persistence. The clearness index persistence (or "smart persistence") forecast uses the last available observation of Kt as the next forecast. This forecast is difficult to improve upon when the cloud cover stays constant or when skies are clear. Considering that our forecast sites are in the Sacramento Valley of California where the majority of the time it is clear, this simple method is expected to perform relatively well and will be difficult to improve upon. When the clearness index is multiplied by the TOA GHI to convert back to the GHI at the surface if operations require, it inherently corrects for changes in solar elevation with time.

5. Cloud regime classification

To test our hypothesis that splitting the data into subsets based on cloud regimes can improve overall GHI forecast accuracy, we classify the cloud regimes with the k-means algorithm before separate ANNs are trained and tested for each cloud regime subset. The k-means algorithm clusters data by separating samples into k groups by minimizing the within-cluster sum-of-squared departures from the cluster mean, hereafter referred to as sum-of-squares. The process begins by dividing a set of samples (N) into k clusters, each of which is described by the mean (centroid) of the cluster's instances. The k-means clustering algorithm selects the optimal centroid for each cluster by finding the centroid with the minimum within-cluster-sum of squares, i.e. to find the minimum of,

$$\sum_{k=1}^{K}\sum_{i=1}^{N}\|x_i - \mu_k\|^2, \tag{1}$$

where the minimization is computed over each instance i in cluster k. All predictors are normalized before being clustered to avoid having the sum-of-squares dominated by the inputs with the largest magnitudes.

We test five different input subsets to determine the best inputs for the k-means algorithm, summarized in Table 2. The goal is to provide the k-means algorithm with inputs that can physically represent the current cloud cover characteristics. The first input set tested only included the past four 15-min clearness index observations. The second input set tested additionally included spatial clearness index information: the previous 15-min average and standard deviation across the other seven sites. The third input set included the previous inputs as well as both the most recent change in clearness index (Kt_Prev15-Kt_Prev30, which is named Kt15 − Kt30) and the slope of Kt over the past hour. In the fourth input set, we used almost entirely derived variables. We still included the last observation (Kt = 15 min), but added to the derived variables from the previous input set with the variability (standard deviation) of the past four 15-min averages. In our final (fifth) input set, we added the spatial cloud cover variability (standard deviation) and the squared mean cloud cover as inputs.

We wish to find the best set of inputs and the value of k that balances the accuracy of assigning each data set to a cluster without over-fitting the number of clusters to the training data. We inspected plots (not shown) of the sum-of-squares by the k-means algorithm. From this analysis, the exact best value of k was unclear; however, these plots indicated that somewhere between three and seven was likely optimal. Therefore, the next step tested the predictive ability of the regime subsets to determine the optimal number of cloud regimes. The results of the sensitivity studies indicated that error decreased as the number of regimes increased from three to seven. Therefore, we select seven as the number of regimes (k) in the k-means regime classification.

We analyzed the regime classification and the corresponding irradiance variability within each regime to examine the physical representation of the regimes by the k-means algorithm; however, the high dimensionality of the data provides only limited insight into the physical relationships. Fig. 4 shows four plots for this analysis. The top left subplot compares the regime classification relationship between two inputs, Kt_Prev15 and Kt_Prev30 for k-means on Input Set 1. This subplot shows distinct relationships in the phase space of two of the inputs, Kt_Prev15 and Kt_Prev30, with cases having similar values of these parameters being assigned to the same regime. This plot also indicates that there is greater spread in the predictor values within each regime for the middle range of Kt values, i.e. the partly cloudy conditions, than when it is either mostly clear (black) or mostly cloudy (purple). The top right subplot compares the regime classification relationship in the phase space of Kt_Prev15 and Kt Variability (standard deviation for the previous hour) for k-means on Input Set 5 (the one selected). The cluster relationship is less obvious in this phase space; however, the less interpretable regime classification patterns for higher numbers of inputs can be largely attributed to the "curse of dimensionality" [9]. This implies that as more predictors are added

Table 2
Test of input sets for the k-means classification of cloud regimes.

| Input set 1	Input set 2	Input set 3	Input set 4	Input set 5
Kt_{Prev60}	Kt_{Prev60}	Kt_{Prev60}	Kt_{Prev15}	Kt_{Prev15}
Kt_{Prev45}	Kt_{Prev45}	Kt_{Prev45}	Nearby Kt_{Prev15}	Nearby Kt_{Prev15}
Kt_{Prev30}	Kt_{Prev30}	Kt_{Prev30}	Nearby Kt_{Prev15} Stdev	Nearby Kt_{Prev15} Stdev
Kt_{Prev15}	Kt_{Prev15}	Kt_{Prev15}	$Kt_{Prev15}-Kt_{Prev30}$	$Kt_{Prev15}-Kt_{Prev30}$
	Nearby Kt_{Prev15}	Nearby Kt_{Prev15}	Kt slope (1-hr)	Kt slope (1-hr)
	Nearby Kt_{Prev15} Stdev	Nearby Kt_{Prev15} Stdev	Kt temporal variability (Stdev 1-hr)	Kt temporal variability (Stdev 1-hr)
		$Kt_{Prev15}-Kt_{Prev30}$		Cloud cover Stdev
		Kt slope (1-hr)		Cloud cover squared
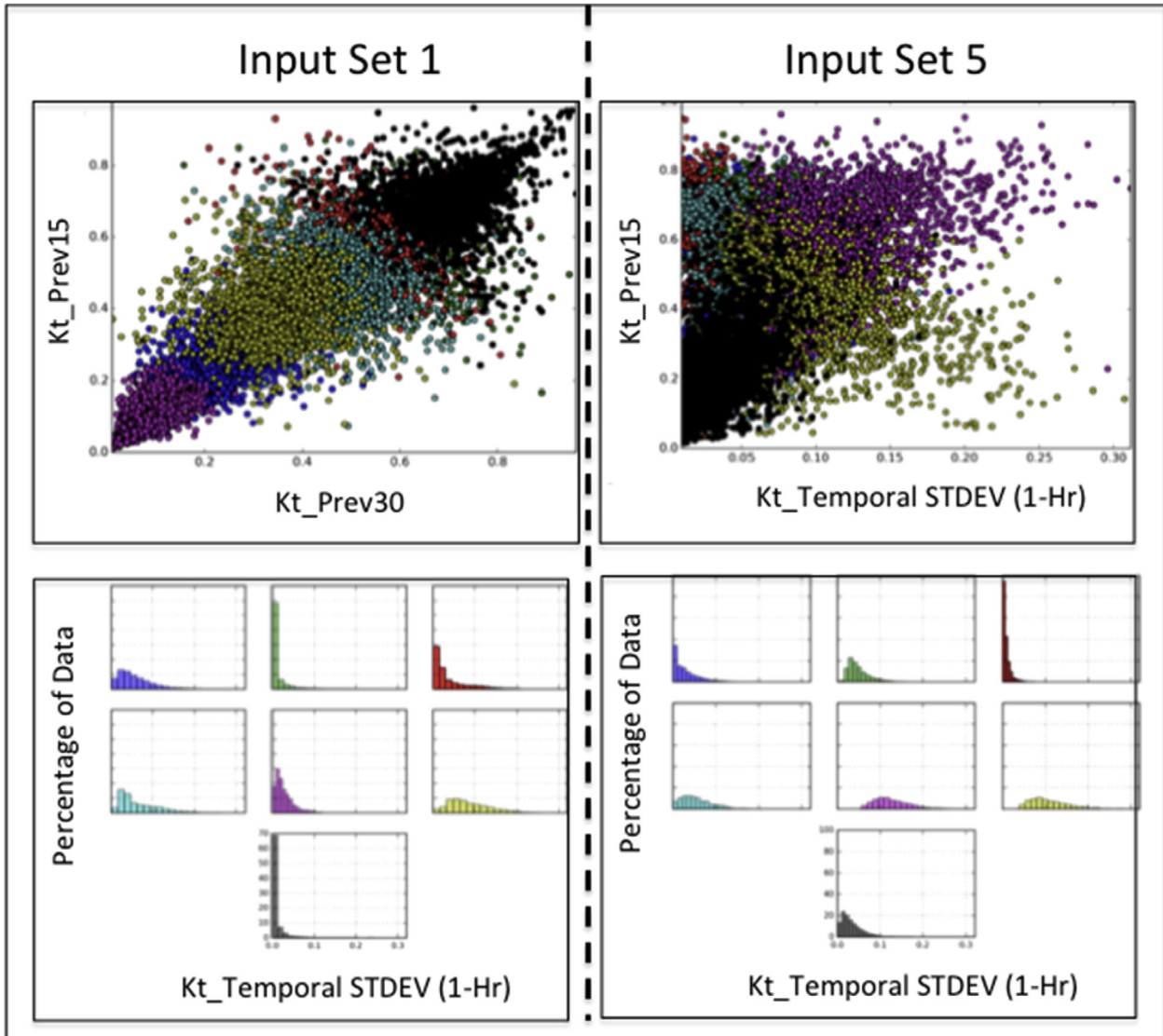
Fig. 4. Analysis of the regime classification (top subplots) for the Input Set 1 (left) and Input Set 5 (right). The analysis for Input Set 1 compares the regime classification for input Kt_Prev15 and input Kt_Prev30. The analysis for Input Set 5 compares the regime classification for input Kt_Prev15 and the Kt_Temporal STDEV (standard deviation previous hour). The bottom subplots are histograms of the Kt_Temporal STDEV for each regime with colors representing the different classification of regimes. These pdfs indicate the difference among the variability across the regimes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to the k-means algorithm, the mapping of clusters in a higher dimensional space results in a loss of interpretability when they are projected onto a two dimensional plane. The bottom left subplot is the 1-hr temporal variability (standard deviation) of the clearness index for each regime as classified by k-means on Input Set 1. The bottom right subplot is the 1-hr temporal variability (standard deviation) of the clearness index for each regime as classified by k-means on Input Set 5. The bottom two subplots highlight an important feature of the k-means regime classification: each regime has a different irradiance variability distribution as is obvious from the varying probability density functions visible in the histograms. This assures us that the k-means algorithm classifies regimes with different underlying irradiance variability distributions because these differing distributions are expected for the various cloud types.

The optimal input subset among those studied for the k-means algorithm was determined to be Input Set 5 based on a sensitivity study evaluated by the MAE for the regime-dependent ANNs
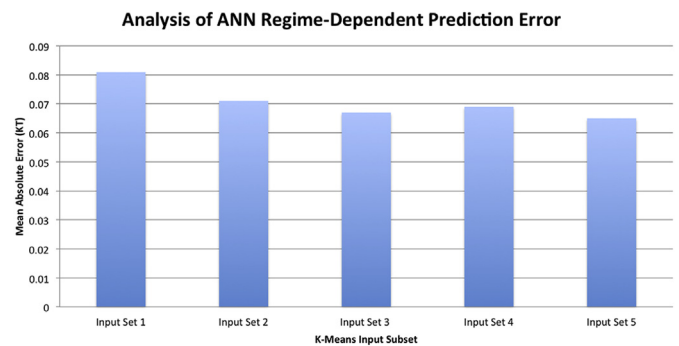

Fig. 5. Sensitivity study of the Regime-Dependent ANNs averaged over all regimes for the 60-min forecast lead time. The regime-dependent ANN on Input Set 5 has the lowest MAE of all input sets tested.

averaged over all regimes for the 60-min forecast lead-time (Fig. 5).

We determine the best input subset as the one with the lowest MAE on the sensitivity test set because we wish to classify cloud regimes that allow the ANNs to model the true relationships among the predictors for each regime.

The results indicate that Input Set 5 has lower error than the other input sets. This is a physically plausible result because Input Set 5 includes the most inputs derived to describe the current cloud pattern. Therefore, the regime identification via k-means is trained and tested with Input Set 5 and seven regimes.

6. Results

6.1. Average results

To analyze the performance of the ANNs trained separately on each of the regimes, the Mean Absolute Error (MAE) for the (in-dependent) testing datasets is computed. The MAE of the ANN prediction for each regime is compared to the MAE for forecasts given by clearness index persistence as well as the MAE given by the ANN trained on all data without regime identification. The MAE is calculated as,

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |(obs(i) - pred(i)|, \qquad (2)$$

where n is the number of instances in the testing data.

The overall results for the clearness index persistence, ANN, and the regime-dependent ANNs show the clearness index persistence method has the lowest error for the 15-min forecast lead time, but the regime-dependent ANN method is best for 60-min, 120-min, and 180-min forecast lead times (Fig. 6). These results highlight the benefit of the regime classification because the regime-dependent ANN method has lower forecast error than the ANN without regime classification at all forecast lead times. It is also important to note that the forecast accuracy improvement over the clearness index persistence increases as the forecast lead-time increases, which is expected since cloud growth and dissipation will lead to larger errors in the clearness index persistence method as lead time increases and also demonstrates the ability of the ANN to predict some of the cloud growth and dissipation.
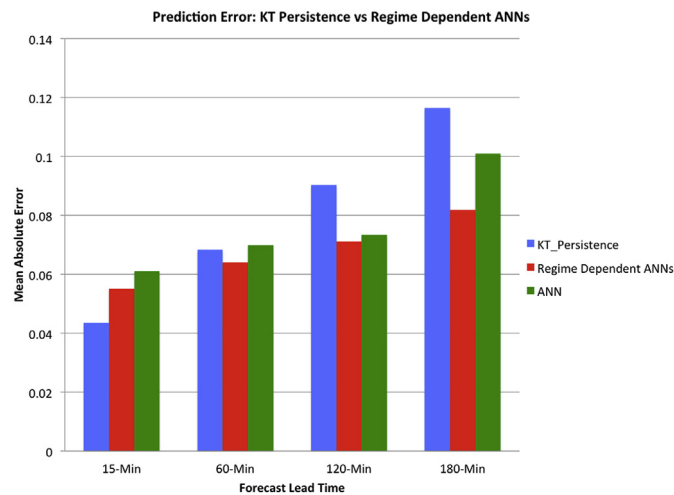

Fig. 6. Comparison of errors for clearness index persistence, ANN, and regime-dependent ANN method at 15-min, 60-min, 120-min and 180-min forecast lead-times. The clearness index persistence is best at 15-min, otherwise the regime-dependent ANN method performs best.

6.2. Regime-dependent results

The percent improvement of the regime-dependent ANN MAE compared to that for the clearness index persistence forecast varies from regime to regime and lead-time to lead-time. This emphasizes that the regime classification allows the ANNs to model the true relationships among the predictors in each regime, which in some regimes produces more forecast error improvement than other regimes due to the complexity of the relationships among the predictors in each subset. The ANN trained without regime classification is also compared to clearness index persistence. For the 15-min lead-time (Fig. 7), neither the ANN nor the regime-dependent ANN is more skillful than clearness index persistence except on the most variable cloud regime (6). In addition to the MAE, we compute the standard deviation of the absolute error to quantify the variability of the forecast error for each regime. This metric is computed as,

$$\sigma_{AE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (|obs(n) - pred(n)| - MAE_j)^2}, \qquad (3)$$

where N is the number of cases in the jth regime. The clearness index persistence MAE for each regime is shown in the second column of Table 3 and the clearness index persistence standard deviation of the absolute error is shown in the third column of Table 3. For regime 6, the MAE of the clearness index persistence is 0.12 and the standard deviation of the absolute error for the clearness index persistence forecast is 0.11. This variability is nearly four times greater than that for Regime 4, for which the standard deviation of the absolute error is 0.03 while the MAE is 0.01.

Such forecast improvements during the most variable regime, i.e. partly cloudy conditions, can aid utility companies and ISOs in planning their units to dispatch. For the forecast lead-times of 60-min (Fig. 8), the regime-dependent ANN reduces MAE compared to that of clearness index persistence in six of the seven regimes. Only for Regime 6 does the ANN trained on all data out-perform the ANN trained on data for the specific regime. For the 120-min forecast lead-time (Fig. 9), the regime-dependent ANN has highest percent improvement in MAE over clearness index persistence in five of the seven regimes. At the 180-min (Fig. 10) forecast lead-time, the regime-dependent ANN is always best. When averaged over all seven regimes, the regime-dependent ANN method reduces the MAE from that of the clearness index persistence forecast by 5.9%, 21.1%, and 29.3% for the 60-min, 120-min, and 180-min forecast lead-times, respectively.
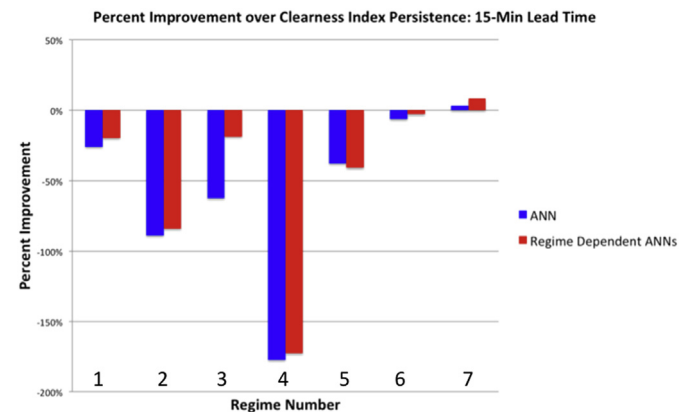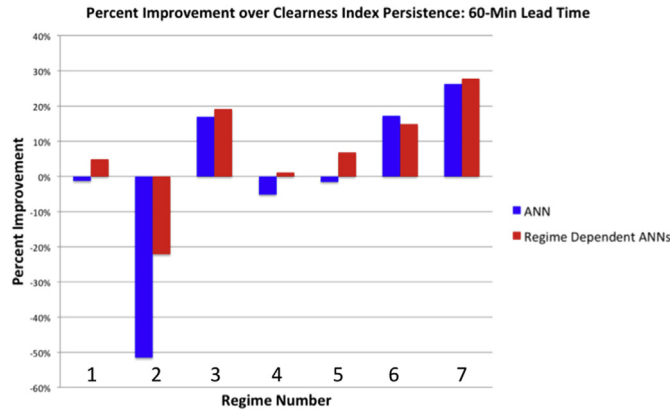

Fig. 7. Percent improvement of the MAE for the ANN and the regime-dependent ANN compared to the clearness index persistence for all seven regimes at the 15-min forecast lead-time.

Table 3
Comparison of the clearness index persistence MAE for each regime to the standard deviation of the absolute error for each regime. The MAEs for each regime are correlated with the variability of the errors, as expected.

	Kt persistence MAE	Standard deviation of Kt persistence MAE
Regime 1	0.10	0.10
Regime 2	0.05	0.05
Regime 3	0.06	0.07
Regime 4	0.01	0.03
Regime 5	0.04	0.07
Regime 6	0.12	0.11
Regime 7	0.06	0.07

Fig. 8. Percent improvement of the MAE for the ANN and the regime-dependent ANN compared to the clearness index persistence for all seven regimes at the 60-min forecast lead-time.
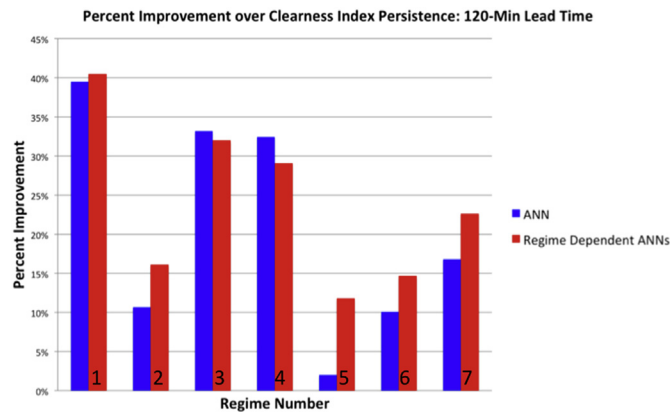
Fig. 9. Percent improvement of the MAE for the ANN and the regime-dependent ANN compared to the clearness index persistence for all seven regimes at the 120-min forecast lead-time.

Fig. 10. Percent improvement of the MAE for the ANN and the regime-dependent ANN compared to the clearness index persistence for all seven regimes at the 180-min forecast lead-time.
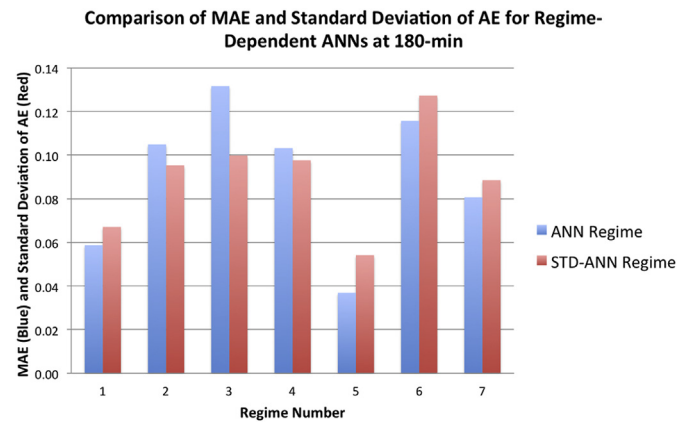
Fig. 11. Comparison of the forecast MAE (blue columns) and the standard deviation of the Absolute Error (red columns) for the regime-dependent ANNs. The regimes with the largest errors correlate with the regimes with the largest standard deviation of the forecast error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 12. Comparison of the forecast MAE and the standard deviation of the Absolute Error for the regime-dependent ANN predictions at all forecast lead times. The regimes with the largest errors correlate with the regimes with the largest standard deviation of the forecast error.

Regimes that are more difficult to predict (i.e. those with variable cloudiness) are expected to have larger forecast uncertainty, which we quantify with the standard deviation of the absolute error. To assess this, the MAE and the standard deviation of the absolute error for the regime-dependent ANNs at the 180-min lead-time are shown in Fig. 11. The plot for the seven regimes demonstrates that different regimes have different average forecast errors and different forecast error variability. The comparison between MAE and standard deviation of the absolute error for all forecast lead times is plotted in Fig. 12. These results show a direct relationship between the magnitude of the forecast error and the variability of the forecast error for all forecast lead times.
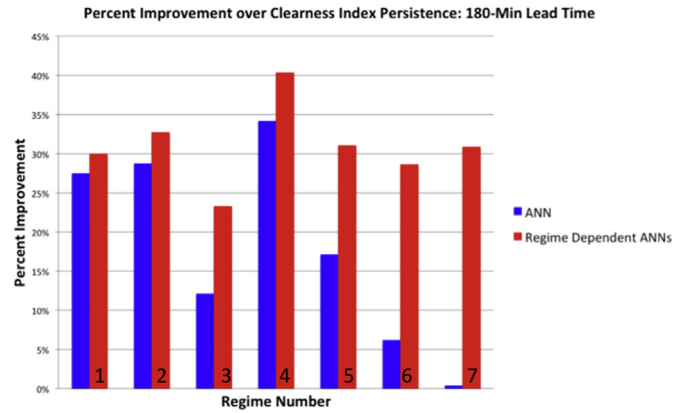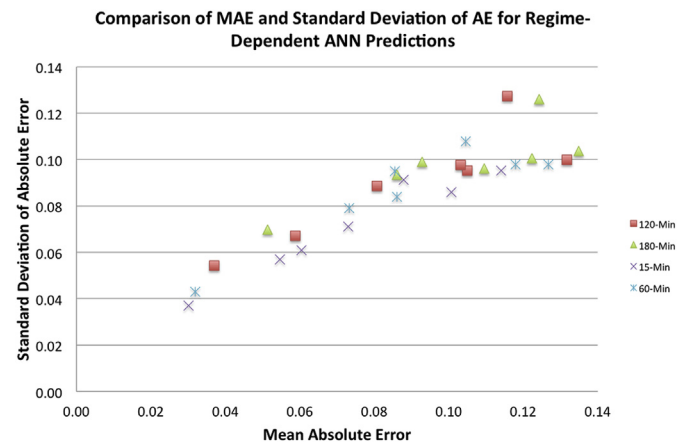
Specifically, the regimes that are more difficult to predict, as

identified by the larger MAEs, also exhibit larger error variability. Therefore, by identifying regimes before applying the ANNs, not only do we increase forecast skill for lead-times of 60-min or more, but also provide a refined estimate of the expected forecast variability.

7. Conclusions and future work

This work has integrated our physical knowledge of the atmosphere with artificial intelligence techniques to demonstrate that solar irradiance forecasts can be improved by recognizing that regime-dependent forecasting improves prediction. We have tested a regime-dependent solar irradiance short-range forecasting system in comparison to a baseline clearness index persistence and a global ANN. The system uses k-means clustering to classify cloud regimes between which the relationship among inputs and solar irradiance is expected to vary. An ANN is then developed for each regime. The results for regime-dependent 15-min average clearness index forecasts, the shortest time frame, show that clearness index persistence forecast nonetheless is more skillful than the new system, with the exception of the regime that has the most variability. For longer lead times (60-min, 120-min, and 180-min), however, the regime-dependent ANNs yield substantial improvement over clearness index persistence. The regime-dependent ANNs also have, on average, lower forecast errors than an ANN trained without regime identification.

In addition to the improvement in forecast accuracy at lead-times of 60-min, 120-min, and 180-min, the regime classification provides value in identifying regimes with greater forecast error variability and thus higher forecast uncertainty. Knowing the forecast error uncertainty aids utility companies in effectively and efficiently managing the power grid.

This paper reports on data from one region, the Sacramento Valley in California. We plan to test these methods for more locations that exhibit different weather regimes, and therefore, different frequencies of cloud types. We also plan to test how satellite data may impact the identification of regimes and the skill of the regime-dependent ANN model.

Acknowledgments

References

[1] O. Bartholomy, T. Vargas, M. Simone, G. Hansen, S. Fitchett, A. Pohl, Benchmarking solar power and irradiance forecasting accuracy at Sacramento municipal utility district, in: IEEE Photovoltaic Specialist Conference, June 8–13, 2014, pp. 63–68.

[2] Y. Chu, H. Pedro, C.F.M. Coimbra, Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning, Sol. Energy 98 (2013) 592–603.

[3] C. Cornaro, F. Bucci, M. Pierro, F. Del Frate, S. Peronaci, A. Taravat, Solar radiation forecast using neural networks for the prediction of grid connected PV plants energy production (DSP Project), in: Proceedings of 28th European Photovoltaic Solar Energy Conference and Exhibition, Sept 30-Oct 4, 2013, pp. 3992–3999.

[4] E. Fernandez, F. Almonacid, N. Sarmah, P. Rodrigo, T.K. Mallick, P. Perez Higueras, 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 (2014) 148–158.

[5] S.J. Greybush, S.E. Haupt, G.S. Young, The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature, Weather Forecast. 23 (2008) 1146–1161.

[6] T.J. Hall, C.N. Mutchler, G.J. Bloy, R.N. Thessin, S.K. Gaffney, J.J. Lareau, Performance of observation-based prediction algorithms for very short-range, probabilistic clear-sky condition forecasting, J. Appl. Meteor. Climatol. 50 (2011) 3–19.

[7] S.E. Haupt, The suncast solar power forecasting system, in: 13th Conference on Artificial Intelligence, Phoenix, AZ, Amer. Meteor. Soc. J6.2, 2015.

[8] M. Hassanzadeh, M. Etezadi-Amoli, M.S. Fadali, Practical approach for sub-hourly and hourly prediction of PV power output, in: North American Power Symposium (NAPS), 1–5, Sept, 2010, pp. 26–28.

[9] M.E. Houle, H.P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Can shared-neighbor distances defeat the curse of dimensionality? scientific and statistical database management, Lect. Notes Comput. Sci. 6187 (2010) 482.

[10] W.P. Mahoney, K. Parks, G. Wiener, L. Yubao, W.L. Myers, S. Juanzhen, L. Delle Monache, T. Hopson, D. Johnson, S.E. Haupt, A wind power forecasting system to optimize grid integration, IEEE Transaction Sustain. Energy 3 (4) (2012) 670–682. Oct. 2012.

[11] T.C. McCandless, S.E. Haupt, G.S. Young, A model tree approach to forecasting solar irradiance variability, Sol. Energy 120 (2015) 514–524.

[12] R. Marquez, C.F.M. Coimbra, Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database, Sol. Energy 85 (5) (2011) 746–756.

[13] R. Marquez, V. Gueorguiev, C.F.M. Coimbra, Forecasting of global horizontal irradiance using sky cover indices, ASME J. Sol. Energy Eng. 135 (2013) 0110171–0110175.

[14] L. Martin, L. Zarzalejo, J. Polo, A. Navarro, R. Marchante, M. Cony, Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning, Sol. Energy 84 (2010) 1772–1781.

[15] A. Mellit, artificial intelligence technique for modeling and forecasting of solar radiation data: a review, Int. J. Artif. Intell. Soft Comput. 1 (1) (2008) 52–76.

[16] A. Mellit, A. Massi Pavan, V. Lughi, Short-term forecasting of power production in a large-scale photovoltaic plant, Sol. Energy 105 (2014) 401–413.

[17] H. Morf, Sunshine and cloud cover prediction based on Markov processes, Sol. Energy 110 (2014) 615–626.

[18] H.T. Pedro, C.F.M. Coimbra, Assessment of forecasting techniques for solar power production with no exogenous inputs, Sol. Energy 86 (7) (2012) 2017–2028.

[19] D.R. Reed, R.J. Marks, Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, MIT Press, Cambridge, MA, USA, 1998.

[20] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386–408.

[21] A. Sfetsos, A.H. Coonick, Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques, Sol. Energy 68 (2) (2000) 169–178.

[22] N. Sharma, P. Sharma, D. Irwin, P. Shenoy, Predicting solar generation from weather forecasts using machine learning, in: Proceedings of the 2nd IEEE International Conference on Smart Grid Communications, Brussels, 17-20 October, 2011, pp. 32–37.

[23] F. Wang, Z. Mi, S. Su, H. Zhao, Short-term solar irradiance forecasting model based on artificial neural network using statistical feature Parameters, Energies 5 (2012) 1355–1370.

[24] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.

[25] D. Yang, P. Jirutitijaroen, W.M. Walsh, Hourly solar irradiance time series forecasting using cloud cover index, Sol. Energy 86 (2012) 3531–3543.