A REGIME-DEPENDENT BAYESIAN APPROACH TO SHORT-TERM SOLAR IRRADIANCE FORECASTS

J6.5

Tyler C. McCandless[1,2*], Sue Ellen Haupt[1,2], George S. Young[2], and Andrew J. Annunzio[1,]

[1]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO

[2]Department of Meteorology, The Pennsylvania State University, University Park, PA

1. INTRODUCTION

The generation of solar power is intermittent due to the advection, growth, and dissipation of clouds. Generators that are dispatched by a utility company or independent systems operators must balance this intermittent source of power. In order to appropriately balance the expected energy demand, utility companies and independent systems operators will increasingly depend upon accurate solar power nowcasting for real-time dispatch of its units.

The prediction of solar power through statistical techniques has gained research interest in the last decade. Sharma et al (2011) found a Support Vector Machine approach produced the lowest Global Horizontal Irradiance (GHI) forecast error. Hassanzadeh et al (2011) and Dazhi et al (2012) found AutoRegressive Integrated Moving Average (ARIMA) models most accurate for short-term predictions of solar power and solar irradiance while Morf (2014) used a Markov process to predict sunshine and cloud cover. Mellit (2008) states that 37 studies have used Artificial Neural Networks (ANNs) in the modeling and prediction of solar radiation. More recently, Martin et al. (2010), Hall et al (2011), Marquez and Coimbra (2011), Wang et al (2012), Chu et al (2013), and Cornaro et al (2013) determined that a final model based on ANNs improves solar irradiance or solar power forecast accuracy.

It is well known that solar irradiance will have varying limits of predictability depending on the forecast site and its weather conditions. Weather regimes with broken clouds will produce a more intermittent source of power from a photovoltaic power plant than weather regimes that are either clear or fully cloudy. Several studies have examined the predictive skill of statistical forecast models in various weather conditions. Pedro and Coimbra (2012) found the accuracy of an ANN optimized with a Genetic Algorithm had a strong seasonal dependence. Marquez et al (2012) correlated total sky images, infrared data, and solar radiation observations at the surface to use as input into an ANN and found the variability of solar radiation to be strongly dependent on clouds. 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 in Mellit et al (2014). Fernandez et al (2014) concluded that the ANN model has accurate performance for days characterized by direct irradiance (clear days) and for days characterized by diffuse irradiance (cloudy days).

This work tests the blending of cloud regime classification and artificial intelligence forecasting to produce a more accurate GHI forecast. Similar to the methodology by Greybush et al (2008), who classify and identify weather regimes before applying optimal weights to ensemble forecasts, we employ a K-Means Clustering technique on various sources of weather and cloud data. ANNs are then implemented on each weather regime independently.

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 (Marquez et al 2013)

We wish to make short term predictions for multiple sites near Sacramento, California for the next 15 minute interval. In operational forecasting, these short-term predictions are blended with forecasts from Numerical Weather Prediction Models and a satellite based cloud advection technique in the National Center for Atmospheric Research SunCast System that predicts solar power out to 168-hours.

Section 2 focuses on our methodology. In section 3, we discuss the datasets: the SMUD irradiance network and the DICast/METAR predictor network. In section 4, we summarize the methods of cloud regime identification. In section 5, the baseline clearness index persistence forecast and the non-linear forecasting technique ANNs are described. In section 6, conclusions are explained and future work is presented.

2. METHODOLOGY

The goal of this work is to develop a regime-dependent short-term prediction of solar radiation. Our methodology begins by identifying the cloud regime, and then employs an Artificial Neural Network to make a prediction for each individual regime as depicted in Figure 1. The first step determines which set of inputs is best for cloud regime classification by the K-Means Clustering algorithm. Then this best set of inputs is used for classification of regimes using a K-Means Clustering algorithm. Finally, ANNs are constructed on each of the cloud regime datasets independently. An ANN is also fit using all data without regime-identification for comparison.

• Corresponding author address: National Center for Atmospheric Research, 3600 Mitchell Lane, Boulder, CO 80303

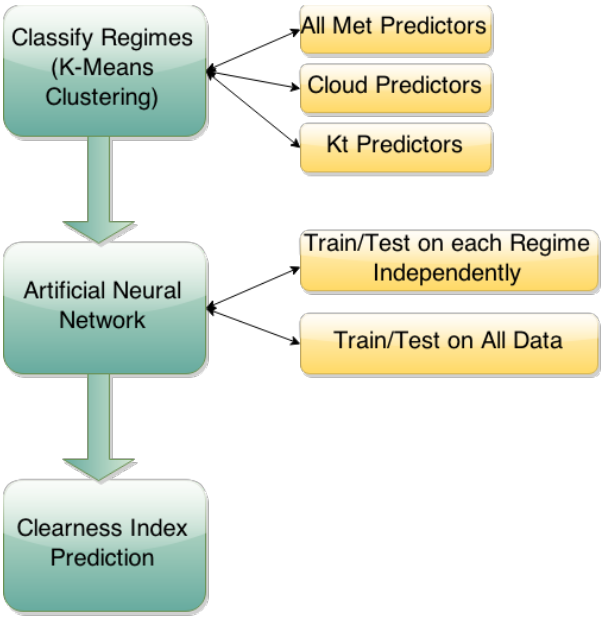

Figure 1. Process design from top to bottom: first classify cloud regimes, then apply ANN models to predict the clearness index.

By identifying the cloud regime before prediction, it is possible to build statistical forecasting techniques specifically for each cloud regime. The statistical learning models, in this case the ANNs, are trained on each cloud regime independently, and thus model each cloud regime. Therefore, in a real-time forecasting environment the predictions are made by identifying the current cloud regime and then applying the model built for that cloud regime to predict the next 15-min average clearness index. The ANNs use weather forecasts and irradiance observations as input to predict clearness index at multiple California locations.

3. DATA

The irradiance observation network used in this study is that of the Sacramento Municipal Utility District (SMUD) in Sacramento, California. We consider data from the eight solar power forecast sites that measure irradiance, shown in Figure 2 as blue triangles. The GHI observations are available from January 25th, 2014 through October 31th, 2014. The temporal resolution of the raw data is one minute and averages are computed over 15 minute intervals. This 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.

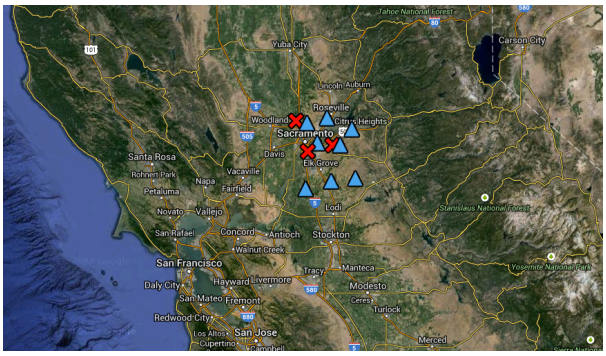

Figure 2. Map of the SMUD sites (blue triangles) and METAR/DICast predictor sites (red X's).

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 were found to be the three locations plotted as red X's in Figure 2. We use six weather variables that are quality controlled and pre-processed with the DICast system: cloud cover, dewpoint temperature, probability of precipitation in the last hour, quantitative precipitation in the last hour, temperature, and wind speed. These six observed weather variables at three stations are combined with the last three 15-min observed clearness index values for a total of 21 predictors. The data are split into 2/3 training and 1/3 testing. We will show our results for the testing dataset.

4. CLOUD REGIME IDENTIFICATION

4.1. *K-Means Clustering*

To test our hypothesis that breaking the training and testing datasets into subsets by cloud regimes can improve overall forecast accuracy, we identify the cloud regime before applying the ANN with the K-Means Clustering algorithm to the individual regimes. The K-Means algorithm clusters data by separating samples into K groups by minimizing the within-cluster sum-of-squares. The process begins by dividing a set of samples (N) into K clusters that are described by the mean of the samples, or centroids, of the cluster. The K-Means clustering algorithm selects the optimal centroid so as to minimize the within-cluster-sum of squares given by,

$$\sum_{i=1}^{N} \min\left(\left|x_j - \mu_j\right|^2\right) \quad , \quad (1)$$

where the minimization is computed for each instance (i and cluster j. The value, $x_j - \mu_j$, is the distance between the vector of instance variables (x_j) and the

cluster center (μ_j). All predictors are normalized before being clustered to avoid having clusters dominated by the predictors with the largest magnitudes.

We examine the percentage of variance explained by the K-Means Clustering algorithm to decide on the best number of clusters (K). The goal is to find the choice of K that best balances the accuracy of assigning each data set to a cluster without over-fitting the number of clusters to the training data. We plot the amount of variance explained as a function of K, to select the optimal number of clusters. Initially the rate of variance explained drops precipitously but the rate decreases as the number of clusters increases. K is chosen at the point at which there is a slowing of the rate of variance by adding more clusters. This analysis was performed for all three predictor sets and the analysis for the clearness index predictors is displayed in Figure 3. Seven clusters were selected due to the slight change in the slope, the average within-cluster sum-of-squares levels off as the number of clusters increases. Similar plots for Dataset A and Dataset B showed five as the best number for K and a summary of the datasets, the predictors and the number of clusters appears in Table 1.

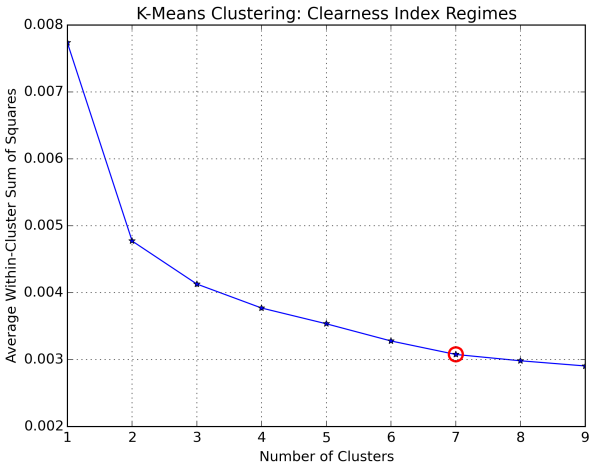

Figure 3. Within-cluster sum-of-squares plotted as a function of the number of clusters. The circle highlights that seven clusters were selected in this case.

Table 1. Description of the three data sets used to determine the best data for clustering cloud regimes.

Name	Predictors	Number of Clusters
All Met Predictors (Dataset A)	• Cloud Cover • Dewpoint Temperature • Probability Precip 1-hr • QPF 1-hr • Temperature • Wind Speed • Clearness Index last 45-min (three 15-min intervals)	5
Cloud Predictors (Dataset B)	• Cloud Cover • Probability Precip 1-hr • Clearness Index last 45-min (three 15-min intervals)	5
Clearness Index (Kt) Predictors (Dataset C)	• Clearness Index last 45-min (three 15-min intervals)	7

We analyzed the clusters selected for each set of input variables by examining groupings by each predictor. The resulting series of plots demonstrate the regime classification for each predictor versus the regime classification for the previous 15-min clearness index value. The previous 15-min clearness index is the last available observation and is the value used for the clearness index persistence forecast. This analysis was accomplished for each of the predictors to examine patterns in the K-Means clustering algorithm.

Figure 4 is an example plot comparing 15-min average clearness prediction to the 15 to 30-min average clearness index observations. This plot reveals some interesting features of the results of the clustering. First, the regime classifications correspond to different magnitudes of the clearness index, which was anticipated. Second, in the middle range of the clearness index values, the clusters are based not only on the values of the clearness index, but also on the trend of the clearness index. For instance, the regime colored black includes instances where the previous 30-min clearness index value is centered around 0.7, while for the previous 15-min clearness index value is centered around 0.5. Thus, this regime exhibits a decreasing trend in clearness index. The opposite is true for the regime colored light blue. This observation indicates that the K-Means algorithm clusters include information regarding the trend in the clearness index

values as well as the clearness index values themselves.

The Dataset C was selected for regime classification because the regime-dependent ANNs error was lowest when using Dataset C for regime classification and the minimal amount of data needed to cluster regimes.

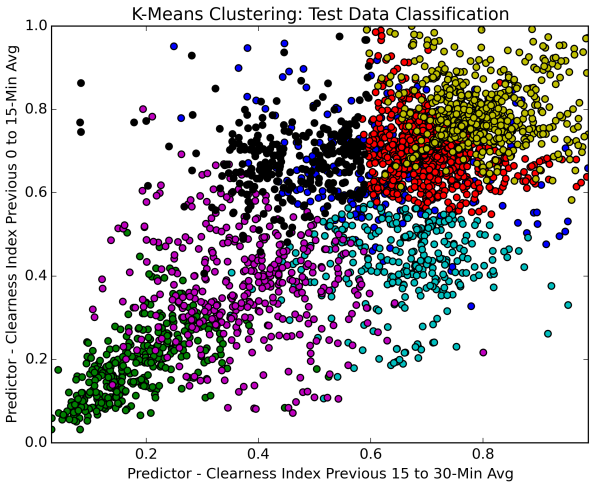

Figure 4. Comparison of the K-Means regime classification for two predictors: the previous two 15-min average clearness index observations for the clearness index dataset. Each color represents a different cluster, or regime, classification.

4.2. *Analysis of Clearness Index Variability*

Regimes with greater temporal clearness index variability (variable cloudiness) are expected to be more difficult to predict. To assess this variability, the standard deviation of the clearness index over the past three 15-min averages was computed. Figure 5 shows clearness index variability histograms for each regime classified on the Dataset C. The colors correspond to the same clusters plotted in Figure 4. The subplots for the seven regimes demonstrate that different regimes have different distributions of clearness index variability and we expect the forecast uncertainty to correspond to the magnitude and distribution of each regime's clearness index variability. The generally clear (green) and cloudy regimes (red and yellow) made a mode near zero while the partly cloudy regimes tend to have a higher mode.

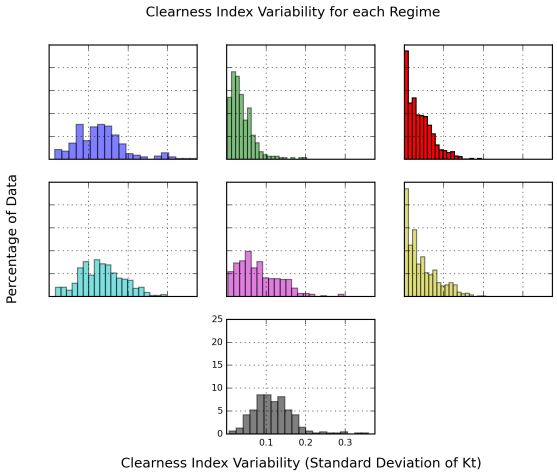

Figure 5. Subplots for each cloud regime's temporal variability for the regimes identified on Dataset C. Each subplot has twenty bins for each cloud regime and each color correlates with the same cluster color from Figure 4.

5. PREDICTION TECHNIQUES

5.1. *Clearness Index Persistence*

We wish to compare our prediction techniques to a baseline prediction method. The clearness index persistence forecast is our chosen baseline. 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, or lack thereof, is constant. When multiplied by the TOA GHI, it inherently corrects for changes in azimuth with time.

5.2. *Artificial Neural Networks*

We employ the ANN as the non-linear Artificial Intelligence (AI) prediction technique for our predictions . ANN's 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 Mellit (2008), the AI models used in many studies have been successfully developed to model solar radiation, clearness index, and insolation with no transformations of the data necessary for prediction. Sfetsos and Coonick (2000) 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.

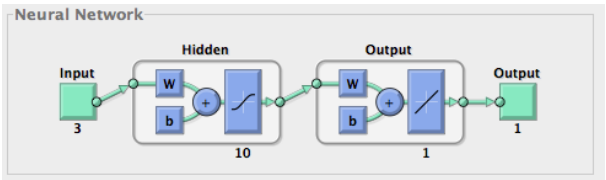

Figure 6. Schematic of a feed-forward Artificial Neural Network used in this study.

The ANN used here (Figure 6) is a feed-forward neural network trained by a backpropagation algorithm, also known as a multi-layer perceptron (Rosenblatt 1958). This ANN configuration has several tunable parameters that were determined from a sensitivity study on a training dataset. The best ANN configuration was selected based on a sensitivity study that produced the lowest test errors for the dataset without regime-identification (not shown). The configuration of the ANN with the lowest error had one hidden layer with ten hidden nodes.

5.3. *Regime-Dependent ANN Prediction Results*

To analyze the performance of the ANNs trained on each of the regimes independently, the Mean Square Error (MSE) on the testing datasets is computed. The MSE of the ANN for each regime is compared to the MSE for forecasts given by clearness index persistence. The MSE is calculated as,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}((obs(i) - pred(i))^2 \quad , \quad (2)$$

where n is the number of instances in the testing data. To compare the skill of the regime-dependent ANNs to the clearness index persistence, the percent improvement of the MSE is computed and recorded in Table 2. Column two displays the percent improvement with each regime's testing dataset size listed in column three and the variability of the clearness index shown in column four. The first important feature of these results is the differences in percentage improvement across the regimes. The percent improvement of the regime-dependent ANN MSEs over that for clearness index persistence varies from -30.9% to 55.3%. The regime with the highest clearness index variability, regime four, has the highest percent improvement over clearness index persistence. Forecast improvements during the most variable regime, i.e. partly cloudy conditions, can aid utility companies and independent systems operators in planning their units to dispatch. The negative percent improvements are likely amendable by regime-dependent tuning of the weight decay term in the ANN configuration to give more weight to the previous clearness index observations, which will be addressed in future research. Finally, the average percent improvement for all regimes is 17.8%, which is a substantial improvement over the clearness index persistence. In addition, the results for the regime-dependent ANNs had lower mean error compared to the ANN that was trained without regime identification.

One can further reduce the overall error by eliminating the cases of negative improvement, effectively what the wait decay mentioned above would achieve. To accomplish this, we use the forecasts made with the regime-dependent ANNs on all regimes except regimes two and seven, which showed negative improvement. For these cases, the predictions were replaces with the clearness index persistence. The resulting combination of the two approaches showed a mean percentage improvement over clearness index persistence increased to 20.7%. It is likely that an increase in the size of the training and testing datasets, as well as fine-tuning the ANN configuration for the regimes with the highest errors will lead to greater improvements over the clearness index persistence.

Table 2. Percent improvement of the MSE for the regime-dependent ANNs compared clearness index persistence in column 2. Columns 3 and 4 describe the size of the testing data for each regime and the average standard deviation of the clearness index over the past three 15-min observations.

Clearness Index Regime	Percent Improvement Over Kt Persistence	Testing Data Size	Kt Variability (standard deviation)
1	10.5%	324	0.12
2	-30.9%	336	0.04
3	8.1%	1216	0.04
4	55.3%	306	0.13
5	47.2%	398	0.09
6	20.9%	796	0.04
7	-24.7%	285	0.11
Average	17.8%		

6. CONCLUSIONS AND FUTURE WORK

We have tested a very short-range regime-dependent solar irradiance prediction system. The system uses K-Means clustering to classify cloud regimes and applies an ANN to each regime independently. The preliminary results for the regime-dependent 15-min average clearness index forecasts show substantial improvement over the baseline clearness index persistence forecasts.

This paper reports on data from one region, the Sacramento Valley in California. We plan to test these methods for more locations and a longer time series. In addition, clearness index forecasts will be made at 15-min intervals out to three hours. A hierarchical Bayesian linear regression technique will be also be tested to create calibrated forecast uncertainty estimates.

Acknowledgements: This research is supported by the Department of Energy Sunshot Program (under award DE-EE0006016) and the National Center for Atmospheric Research. The authors wish to thank the entire SunCast Project Team Members as well as Tyler's PhD committee members for helpful guidance and suggestions. Thanks also go to Sacramento Municipal Utility District for data access and to Tom Brummet of NCAR for data collection and processing.

References

Chu, Y., H. Pedro, and C.F.M. Coimbra, 2013: Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning. *Solar Energy*, **98**, 592-603.

Cornaro, C., F. Bucci, M. Pierro, F. Del Frate, S. Peronaci, and A. Taravat, 2013: Solar Radiation Forecast Using Neural Networks for the Prediction of Grid Connected PV Plants Energy Production (DSP Project), *Proceedings of 28th European Photovoltaic Solar Energy Conference and Exhibition*, Sept 30 - Oct 4, 3992 – 3999.

Marquez, R., V. Gueorguiev, and C.F.M. Coimbra (2013) "Forecasting of Global Horizontal Irradiance Using Sky Cover Indices, *ASME Journal of Solar Energy Engineering,* **135**, 0110171-0110175.

Fernandez, E., F. Almonacid, N. Sarmah, P. Rodrigo, T.K. Mallick, and P Perez-Higueras, 2014: A model based on artificial neuronal network for the prediction of the maximum power of a low concentration photovoltaic module for building integration. *Solar Energy*, **100**, 148-158.

Fu, C-L., and H-Y. Cheng, 2013: Predicting solar irradiance with all-sky image features via regression. *Solar Energy,* **97**, 537-550.

Greybush, S.J., S.E. Haupt, and G.S. Young, 2008: The Regime Dependence of Optimally Weighted Ensemble Model Consensus Forecasts of Surface Temperature. *Wea. Forecasting*, **23**, 1146–1161.

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

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

Marquez, R., and C.F.M. Coimbra, 2011: Forecasting of Global and Direct Solar Irradiance Using Stochastic Learning Methods, Ground Experiments and the NWS Database, *Solar Energy,* **85:5**, 746-756.

Marquez, R., V. Gueorguiev, and C.F.M. Coimbra, 2013: "Forecasting of Global Horizontal Irradiance Using Sky Cover Indices, *ASME Journal of Solar Energy Engineering,* **135**, 0110171-0110175.

Martin, L., Zarzalejo, L., Polo, J., Navarro, A., Marchante, R., and Cony, M, 2010: Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning. *Solar Energy*, **84**, 1772-1781.

Mellit, A., 2008: Artificial Intelligence Technique for Modeling and Forecasting of Solar Radiation Data: A Review. *Int. Journal Artificial Intelligence and Soft Computing*, **1:1**, 52-76.

Mellit, A., Massi Pavan, A., and V. Lughi, 2014: Short-Term Forecasting of Power Production in a Large-Scale Photovoltaic Plant, *Solar Energy*, **105**, 401-413.

Morf, H., 2014: Sunshine and Cloud Cover Prediction Based on Markov Processes, *Solar Energy*, **110**, 615-626.

Pedro, H. T., and C.F.M. Coimbra, 2012: Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, ***86*:7**, 2017-2028.

Rosenblatt, F., 1958: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *In. Psychological Review*, **65:6**, 386-408.

Sharma, N., P. Sharma, D. Irwin, and P. Shenoy, 2011: Predicting Solar Generation from Weather Forecasts Using Machine Learning: *Proceedings of the 2nd IEEE International Conference on Smart Grid Communications*, Brussels, 17-20 October = pp. 32-37.

Wang, F., Z. Mi, S. Su, and H. Zhao, 2012: Short-Term Solar Irradiance Forecasting Model Based on Artificial Neural Network using Statistical Feature Parameters, *Energies*, **5**, 1355-1370.

Yang, D., P. Jirutitijaroen, and W. M. Walsh, 2012: Hourly Solar Irradiance Time Series Forecasting Using Cloud Cover Index, *Solar Energy*, **86**, 3531-3543.