Hindawi
Journal of Renewable Energy
Volume 2017, Article ID 2437387, 7 pages
https://doi.org/10.1155/2017/2437387

*Research Article*

# Feature Selection and ANN Solar Power Prediction

**Daniel O'Leary[1] and Joel Kubby[2]**

[1]*Department of Computer Science, City College of San Francisco (CCSF), Mailbox LB8, 50 Phelan Ave., San Francisco, CA 94112, USA*
[2]*Jack Baskin School of Engineering, University of California, Santa Cruz, 1156 High Street, Mail Stop SOE2, Santa Cruz, CA 95064, USA*

Correspondence should be addressed to Joel Kubby; jkubby@soe.ucsc.edu

Received 8 May 2017; Revised 14 September 2017; Accepted 16 October 2017; Published 8 November 2017

Academic Editor: Ben Xu



A novel method of solar power forecasting for individuals and small businesses is developed in this paper based on machine learning, image processing, and acoustic classification techniques. Increases in the production of solar power at the consumer level require automated forecasting systems to minimize loss, cost, and environmental impact for homes and businesses that produce and consume power (prosumers). These new participants in the energy market, prosumers, require new artificial neural network (ANN) performance tuning techniques to create accurate ANN forecasts. Input masking, an ANN tuning technique developed for acoustic signal classification and image edge detection, is applied to prosumer solar data to improve prosumer forecast accuracy over traditional macrogrid ANN performance tuning techniques. ANN inputs tailor time-of-day masking based on error clustering in the time domain. Results show an improvement in prediction to target correlation, the $R^2$ value, lowering inaccuracy of sample predictions by 14.4%, with corresponding drops in mean average error of 5.37% and root mean squared error of 6.83%.

## 1. Introduction

Power service providers are increasing the use of solar power due to (among many factors) decreases in the cost of solar power production systems, increases in the cost of traditional energy sources, environmental concerns, and legislative requirements. These same forces increase the prevalence of homes and small businesses with solar panels and storage that produce solar power to use in the home or business or store in battery banks or smart appliances or sell power back to power companies in tiered or real-time pricing structures.

Given solar power's variable, intermittent, and nondispatchable nature, considerable effort has been made to develop accurate forecasts that meet the needs of macrogrid power providers. Forecasting the production of large solar arrays and wind farms allows power providers the time necessary to make changes to base load power plant production to minimize peak power plant use. These forecasts often use artificial neural networks (ANNs) which access multiple and varied data sources to estimate power changes hours or days in advance.

With the rise of variable real-time pricing available to the consumer, prosumers can also benefit from forecasting solar power production, to optimize decisions about power storage, use, and sale. However, prosumers have different datasets, power profiles, and forecasting needs than power providers. For example, prosumers do not have access to cloud motion vector data and they do not need forecasts days in advance, which are required for base power plants to achieve steady-state power output. ANNs are strongly dependent on scale, resolution, and forecast variables [1]; the ANNs developed for power companies are not suitable for prosumer use. Prosumers require short-term predictions based on limited data sources. The ANN investigated is intended for use at the prosumer level, which is less likely to have access to complex weather data; therefore, this model uses measurements of variability in the irradiance measurements to assess cloud cover. Further, this model's focus on the prosumer needs and the mitigation of the nondispatchable nature of renewable energy have dictated this model's concentration on short-term forecasting. With accurate solar power prediction, prosumers can make decisions about storing, using, or selling power based on reasonable future

expectation and maximizing the value returned from a solar investment.

ANN's success is strongly correlated to careful parameter tuning. Input masking, a parameter tuning technique used in visual recognition systems, has been applied successfully in applications for audio signal classification and wind turbine power forecasting. This paper details the use of input masking in ANNs unique to short-term solar power forecasting.

## 2. Related Literature and Motivation

In 2005, Mellit et al. [2] used a radial basis function (RBF) based artificial neural network (ANN) to predict daily global solar radiation. Using 20 years of global solar radiation data from a meteorological station in Algiers, researchers created a composite reference year. Using this reference year for training, validation, and test, the scientists were able to create an ANN with one hidden layer of nine nodes that took inputs of air temperature and sunshine duration and output global daily solar radiation. According to the article, these inputs' relationship to the output is nonlinear and poorly suited to other nonlinear signal predictors; however, the accuracy of these other methods is not discussed, so it is unknown to what degree the RBF is an improvement.

Solar power production forecasts on large solar farms with no exogenous inputs [3] have been used to compare different forecasting models. One- and two-hour forecasts found that ANNs outperformed other techniques in terms of mean absolute error (MAE) and mean bias error (MBE).

Researchers have used backpropagation of ANNs for short-term PV power generation prediction [4] in Istanbul, Turkey. Using inputs of ambient temperature, cell temperature, diffuse solar irradiation, and power produced, a comparison was made of the increase in root mean squared error (RMSE) and correlation coefficients for varying time horizons of the prediction. The authors found that a 5-minute time horizon held strong prediction accuracy, and anything between 5 and 35 minutes had an acceptable RMSE and correlation coefficient. The authors also did not identify what is considered an acceptable correlation coefficient or RMSE, but based on the charts in the paper, it would appear to be a correlation coefficient of 0.85 or greater and an RMSE of 50 W (from a 750 W panel) or less. Machine learning techniques are increasingly used for solar power prediction, including researchers in Australia who used machine learning techniques for the prediction of power based on past power production and weather readings [5] and researchers in Houston, TX, who demonstrated the value of machine learning predictions of solar power in urban microgrids based on humidity and time of day [6].

Power prediction is primarily the domain of mid to large grid power producers; however, power prediction has proven valuable in rural areas without grid tied electrical systems. Researchers developed an energy management system (EMS) for a microgrid in Huatacondo, Chile, a small village in the isolated Atacama Desert [7]. Using a backpropagation artificial neural network, predictions on power consumption had a mean absolute error (MAE) of 1.6 kW and a standard deviation of 1.4 kW. The percentage of error associated with these values is not given in the paper, but based on the graphs, the power consumption described, and the given MAE, a 10% error is a reasonable estimate.

Table 1: A glossary of variables used in ANN error equations.

| Variable | Definition |
|---|---|
| $m$ | The number of samples in the evaluation |
| $t$ | The sample index |
| $P_t$ | Power produced at time $t$ |
| $\widehat{P}_t$ | The forecasted power for time $t$ |
| $\overline{P_t}$ | Average power = $(1/m)\sum_{t=1}^{m} P_t$ |

The masking of ANN inputs to reduce error has been developed successfully for edge detection in image processing systems. These techniques have been applied in bioacoustic signal detection [8] and wavelet analysis in wind farm power production forecasts [9].

## 3. Assessing ANN Accuracy

Several benchmarks are used to compare the quality of different ANN and machine learning tools. Table 1 defines the terms used in the ANN assessments.

Common error assessment metrics are as follows.

(i) Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{m}\sum_{t=1}^{m}\left|P_t - \widehat{P}_t\right|. \tag{1}$$

(ii) Mean bias error (MBE):

$$\text{MBE} = \frac{1}{m}\sum_{t=1}^{m} P_t - \widehat{P}_t. \tag{2}$$

(iii) Root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}\left(P_t - \widehat{P}_t\right)^2}. \tag{3}$$

(iv) Normalized root mean squared error (nRMSE):

$$\text{nRMSE} = \sqrt{\frac{\sum_{t=1}^{m}\left(P_t - \widehat{P}_t\right)^2}{\sum_{t=1}^{m} P^2}}. \tag{4}$$

(v) Correlation coefficient ($R$):

$$R^2 = 1 - \frac{\sum_{t=1}^{m}\left(P_t - \widehat{P}_t\right)^2}{\sum_{t=1}^{m}\left(P_t - \overline{P_t}\right)^2}. \tag{5}$$

Table 2 consists of the benchmarks outlined in this section and the corresponding measurements for ANN accuracy found in the related literature in Section 2.

Table 2: ANN assessments from the related literature.

| Assessment | Related literature values |
|---|---|
| MRE | 1.5% [3] |
| $R^2$ | 95% [3], 85% [4] |
| MAE | 53.49 kW [3], 10% [7] |
| nRMSE | 15.82% [3], 7% [4] |

## 4. Data

This work uses data that was collected at five-minute intervals continuously from May 2011 to August 2012. The data was collected from measurements on a dual axis tracking polycrystalline silicon photovoltaic module with 170-watt maximum power installed at the RE Lab at NASA Ames in the Moffett Field Air Force Base in Mountain View, California. This data and the corresponding predictions include night time readings, when energy production is zero. Including night data allows us to use the full day to identify time regions of high and low prediction accuracy for input masking. The data samples consist of four measurements:

(i) *Timestamp*. The date and time when samples are taken.

(ii) *Normal Incidence Pyrheliometer (NIP)*. A measure of the direct beam solar irradiance ($W/m^2$).

(iii) *Precision Spectral Pyranometer (PSP)*. A measure of the total irradiance in the plane of the array ($W/m^2$).

(iv) *Maximum Power Point (MPP)*. A measure of the power produced from the solar panel (W).

To describe the inputs to the ANN calculated from the measurements above, allow $t_n$ to be the time of a particular sample. Further, let $n$ be an index of the sample taken in relation to the sample under consideration. Thus, $n$ is 0 for the sample in the dataset under investigation and $t_1$ is the timestamp of the next sample taken, five minutes later, and $t_x$ is the $x$th sample taken $5x$ minutes after time $t_0$. Similarly, $t_{-1}$ is the sample taken 5 minutes prior to $t_0$ and $t_{-24}$ is the sample taken 2 hours earlier.

Further, allow the function $P_{\text{MPP}}(t)$ to map time values to MPP samples. $P_{\text{MPP}}(t_0)$ is the power produced for the sample under consideration, $P_{\text{MPP}}(t_{-4})$ is the MPP value for the sample twenty minutes prior to $t_0$, and $P_{\text{MPP}}(t_4)$ is the MPP value for the sample twenty minutes after $t_0$. Similarly, allow $I_{\text{PSP}}(t)$ to be the irradiance measured by the PSP (total plane of array irradiance) at time $t$ and $I_{\text{NIP}}(t)$ to be the irradiance measured by the NIP (direct normal irradiance) at time $t$.

Beyond the measurements themselves, the absolute value of the slope of the MPP curve is calculated using

$$S(x) = \left| \frac{P_{\text{MPP}}(t_0) - P_{\text{MPP}}(t_x)}{t_0 - t_x} \right|. \quad (6)$$

Similar equations are used to find the magnitude of the slope of the PSP and NIP. The magnitude of the slopes of measurement values gives an indication of variably cloudy days. As clouds pass over the irradiation measurement tools, irradiance varies greatly. During sunny days, irradiance slopes remain low.

Table 3: A table defining the variables used as inputs to the ANN.

| Variable | Definition |
|---|---|
| $t_0$ | Current time |
| $P_{\text{MPP}}(t_0)$ | Current MPP |
| $I_{\text{PSP}}(t_0)$ | Current PSP |
| $I_{\text{NIP}}(t_0)$ | Current NIP |
| $P_{\text{MPP}}(t_{-1}) \cdots P_{\text{MPP}}(t_{-24})$ | Past two hours of MPP values |
| $I_{\text{PSP}}(t_{-1}) \cdots I_{\text{PSP}}(t_{-24})$ | Past two hours of PSP values |
| $I_{\text{NIP}}(t_{-1}) \cdots I_{\text{NIP}}(t_{-24})$ | Past two hours of NIP values |
| $S(-1)_{\text{MPP}} \cdots S(-2)_{\text{MPP}}$ | Slope of MPP at five and ten minutes |
| $S(-1)_{\text{PSP}}$, $S(-2)_{\text{PSP}}$ | Slope of PSP at five and ten minutes |
| $S(-1)_{\text{NIP}}$, $S(-2)_{\text{NIP}}$ | Slope of NIP at five and ten minutes |

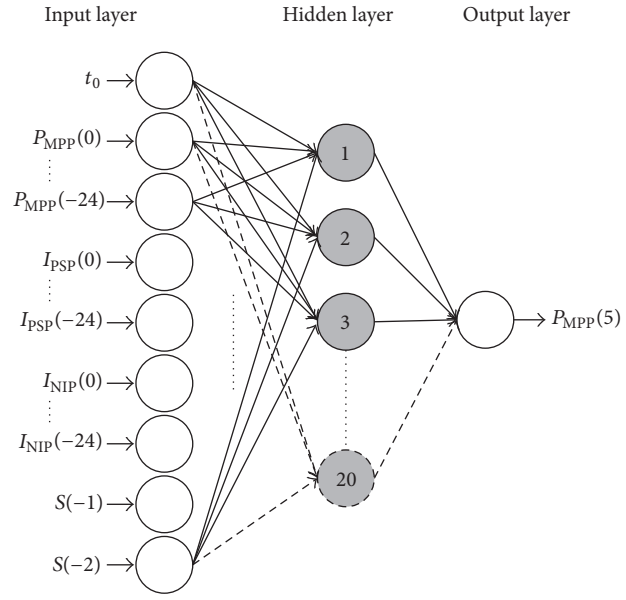


Figure 1: The artificial neural network (ANN) with standard preprocessing contains two previous hours' worth of measurements and two slopes as input and a twenty-minute prediction window.

Knowing this nomenclature, we can now describe the inputs to the ANN in Table 3. Colloquially, each sample includes the current measurements, the power for the previous two hours, and the slope of the previous two samples.

## 5. ANNs

Artificial neural networks are mathematical constructs based on the physical structures of a biological system, the brain. Typically, a preprocessing step is performed on data before entering into the ANN. We have performed a min–max normalization, resulting in all inputs, outputs, and targets to fall between zero and one.

The artificial neural network (ANN) is made up of layers of neurons. Figure 1 depicts the version of the ANN structure

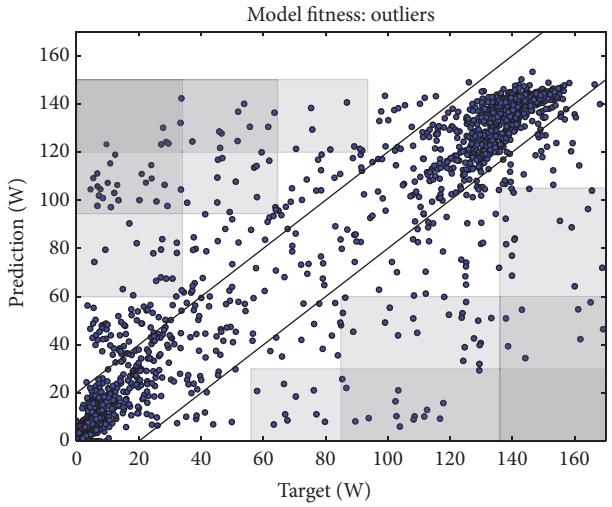


Figure 2: A model fitness graph for Max Power Point (MPP) target versus prediction analysis of a standard, normalized input ANN. Black lines and grey boxes have been added to the diagram to identify and assess target/prediction outliers with high error values.

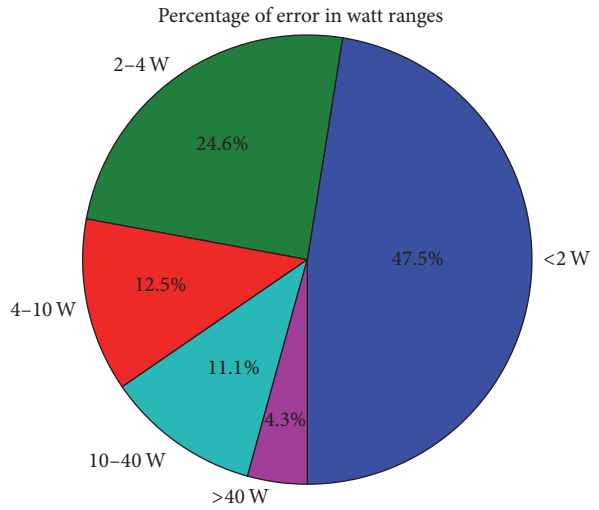


Figure 3: A pie chart of the proportion of predictions in different error categories. Nearly half (47.5%) of all predictions have an error of less than 2 [W].

used with the RE Lab data. This ANN consists of three layers: the input layer, the hidden layer, and the output layer. Each neuron in the input layer takes in one data source. The output of each input layer neuron is input for each of the hidden layer neurons. This ANN has seventy-eight input neurons; each neuron in the hidden layer will have seventy-eight inputs. This ANN relationship occurs a second time between the output of a neuron in the hidden layer and the inputs of the output layer. Thus, if you have twenty neurons in the hidden layer, you will have twenty inputs for each neuron in the output layer.

The ANN model randomly divides the total set of data from the RE Laboratory into three main categories: training, validation, and testing. The test consists of 20% of the total dataset. Of the remaining 80% of the data, 85% is allocated to training and 15% is allocated to validation. The training of an ANN involves multiple cycles of training, referred to as epochs. Each epoch consists of the ANN training on the training dataset and the resulting neural network is then applied to the validation dataset. If the RMSE of the prediction on the validation set is lower than the previous validation RMSE, indicating that the training of the ANN has improved accuracy, training continues. The ANN will continue to train and validate in a cycle until the validation dataset RMSE does not improve for ten consecutive epochs. While the ANN learns from the training dataset, it uses backpropagation to adjust the weights on each neuron based on the error of the output layer and each neuron's output is dictated by a sigmoidal activation function.

## 6. ANNs with Standard Preprocessing

Prior to inputting the RE Lab data into the ANN, measurements were normalized, such that each input varies from 0 to 1. Figure 2 shows the correlation between the predictions of the ANN and our total dataset that was divided randomly into three categories: test, validation, and training. The unmasked ANN was trained on the training data and validated on the validation data, and when the validation showed that the training was done, the resulting ANN was tested on the test data. We then returned to the original, total dataset, divided it randomly into three categories (training, test, and validation), and developed the masked ANN on the training data; once it was validated that the training was done, it was tested on the testing dataset. That is to say, each ANN pulls from the same dataset and uses the same ratio of training : validation : test; however, the contents of each category are unique for each run of each ANN. Figure 2 correlates the predictions of the ANN on the test dataset to the max power point (MPP) of the solar panel 20 minutes in the future (the prediction value, found along the vertical axis) and the actual MPP that was measured twenty minutes later (the target value, found along the horizontal axis). The accuracy measurements for the ANN, with an $R^2$ value of 90.88%, an RMSE of 16.98 W, and a MAE of 6.33 W (for a 170 W panel), are in line with other ANN forecasting accuracy correlation coefficients under similar circumstances [3].

A perfect prediction in the model fitness chart occurs along the 45-degree line. Two black lines have been added to the model fitness chart to indicate a twenty-watt error range, roughly 10%. Further, grey boxes have been added to indicate regions of high error with similar characteristics.

A cursory look at Figure 2 could easily lead the reader to the conclusion that the ANN has a high forecasting error, despite the strong correlation coefficient. However, the model fitness diagram points overlap in areas where the prediction/target points cluster, at low and high power production times. Consequently, the vast majority (ninety percent of points) fall in between the two black lines, indicating a target/prediction pair with an absolute error of less than twenty watts. The pie chart in Figure 3 breaks out the percentage of samples in different absolute error ranges. Less

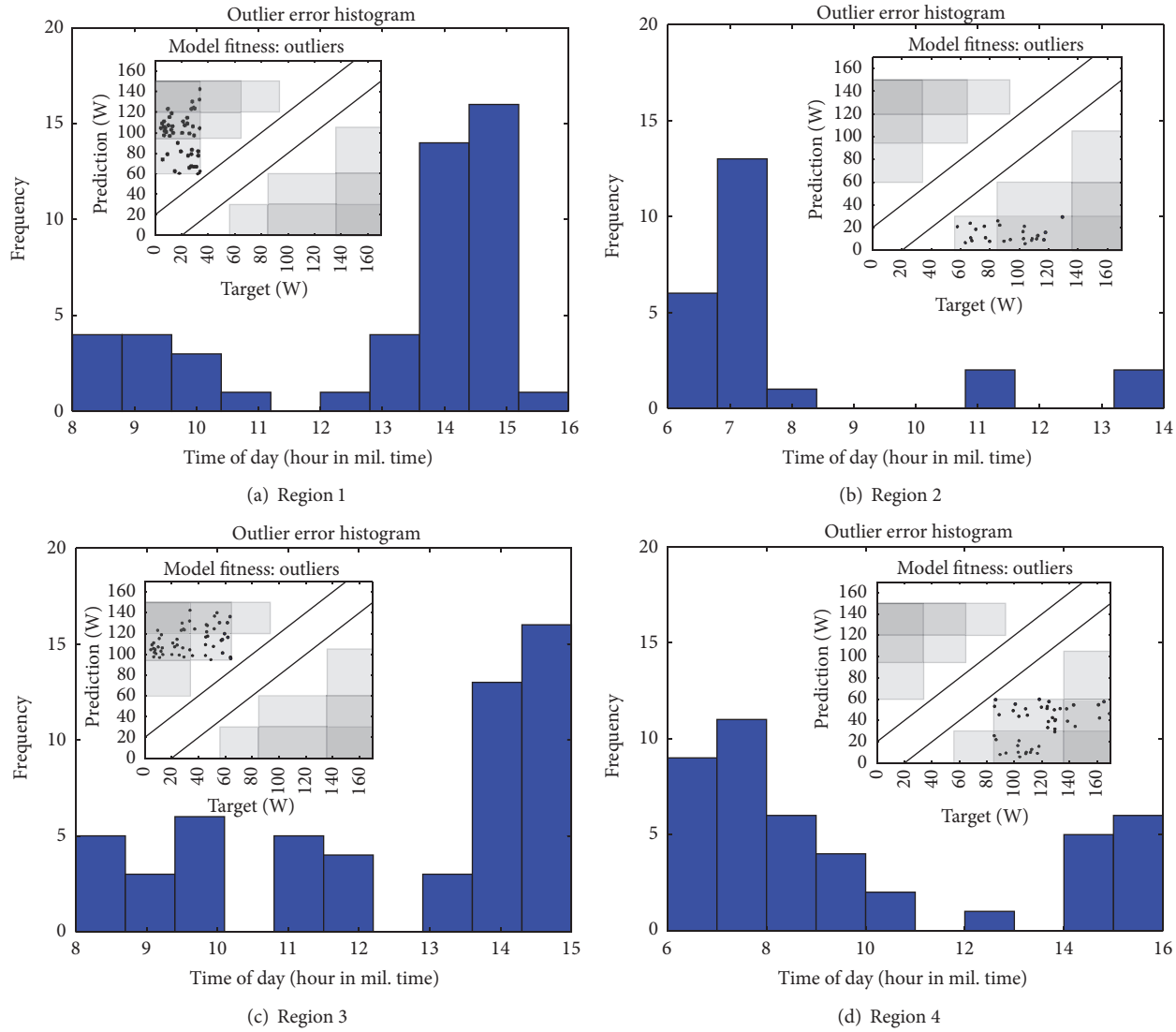


FIGURE 4: Outliers are filtered by region on the model fitness graph. The model fitness graph is reproduced in the inlay of each chart, with the points within the region indicating the area of the graph the region contains. The histograms show the frequency of each outlier in that region grouped by hour of day.

than fifteen percent of all predictions are off by more than ten watts.

A closer examination of the high error forecasts shows a clustering of high error points around specific times of the day. Dividing the model fitness outliers into groups based on their location in the model fitness graph yields a correlation between high error and time of day. Figure 4 examines different regions and their corresponding time-of-day frequency. Figures 4(a) and 4(c) identify regions where the ANN predicts more than it should have. The histograms in Figures 4(a) and 4(c) show the frequency of these forecasts throughout the day, while the inlaid model fitness graph indicates the region being considered. Peaks in the fourteen- and fifteen-hour range (2:00 p.m. to 4:00 p.m.) indicate that these overpredictions are consistent with the time that the panel begins to lose light due to the setting sun and local obstructions. Similarly, the inset model fitness graphs in Figures 4(b) and 4(d) define underpredictions that when reviewed by time of day in the histogram show a high frequency of these errors occurring in the six- and seven-hour range (6:00 a.m. to 8:00 a.m.), corresponding to the time the solar panel first begins to get light in the mornings.

*6.1. Masking Inputs.* Analysis of the standard preprocessing ANN shows four distinct time frames, shown in Figure 5, characterized by the error rates in Figure 4:

(i) Night: when solar energy production is essentially zero.

(ii) Sunrise: one of the two time zones with the highest error rate due to the high volatility of the solar energy production data.

(iii) Day: when solar energy is consistent (on sunny days) and therefore more predictable than sunrise or sunset.

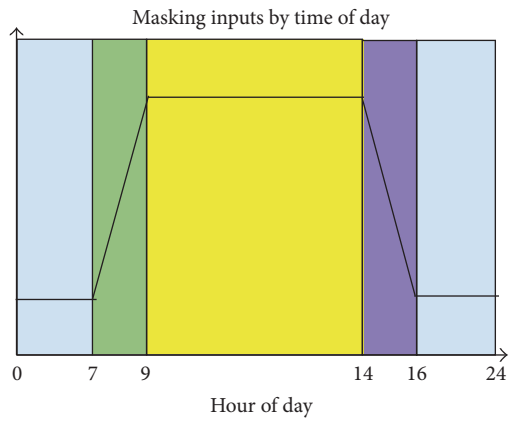


Figure 5: The four time frames of the day, selected by analysis of the cumulative error reviewed by hour. Hour 0 is midnight; hour 12 is noon.

(iv) Sunset: one of the two time zones with the highest error rate due to the high volatility of the solar energy production data.

## 7. ANNs with Enhanced Preprocessing

It should be noted that our use of the terms "sunrise" and "sunset" is not referencing the time of day that the sun encounters the horizon, but the time of day that the sun encounters our instruments; it is sunrise and sunset *from the perspective of our instruments*. In a perfectly flat, empty landscape, these would be the same time, but obstructions, such as trees and buildings, can delay sunrise or hasten sunset from the perspective of our solar panels and irradiance measurement devices.

Using these time frames as the basis for masking the ANN inputs, we replace the PSP input with four inputs: PSP_night, PSP_dawn, PSP_day, and PSP_dusk. PSP_dawn is the same as the original PSP for time frames between 6:00 a.m. and 8:00 a.m., but 0 for all other time frames. Similarly, PSP_day has the same values as the original PSP for time frames between 8:00 a.m. and 2:00 p.m., but it is 0 for all other time frames. The other masked PSP values are created in a similar manner, such that the sum of all masked PSP values would result in the original PSP dataset, as shown in Figure 6.

The same masking procedure is performed for MPP and NIP values. Consequently, the original ANN had 3 inputs at time zero: PSP, NIP, and MPP; the resulting masked ANN has 12 inputs, a night, dawn, day, and dusk for each of the original ANN inputs of PSP, NIP, and MPP. Using the full 24-hour range of data allows the ANN with standard preprocessing to highlight the regions of high variability without knowledge of sunrise and sunset times, making the model robust enough to handle geographic variables such as obstructions and to tailor the model to the unique location of any solar setup. Further, this creates the equivalent of four individual ANN models for each time region in one ANN, allowing the model to have a unique structure for high volatility time frames (and improve accuracy of the least accurate time frames) while also allowing us to make a clear comparison between the unmasked and masked ANNs.

Table 4: Error values for nonmasked versus masked ANN.

| Assessment | Nonmasked | Masked |
|---|---|---|
| $R^2$ | 90.88% | 92.2% |
| RMSE | 16.98 W | 15.82 W |
| MAE | 6.33 W | 5.99 W |

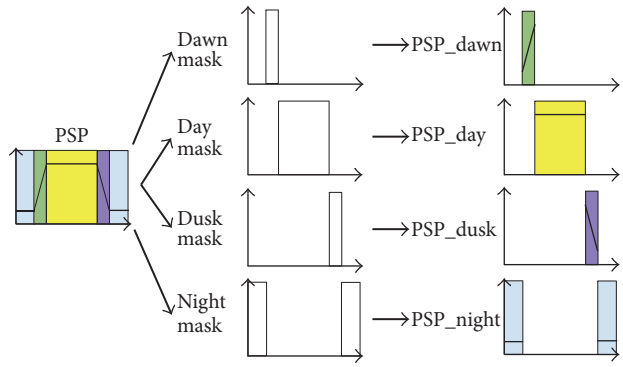


Figure 6: The impact of masking on the PSP inputs. The same masking was applied to the NIP and MPP inputs.

*7.1. Implementing Masked Inputs in the ANN.* The accuracy measurements for the masked ANN, with an $R^2$ value of 92.2%, an RMSE of 15.82 W, and a *MAE* of 5.99 W, show a marked improvement over the nonmasked ANN. The nonmasked $R^2$ value, with an accuracy of 90.88%, offers a potential of 9.12% improvement to be made. The masked $R^2$ value improves this accuracy to 92.2%, an improvement of 1.34%, which in the context of the potential 9.12% improvement is fourteen percent closer to perfect forecasts, and the RMSE and MAE have both dropped accordingly, as outlined in Table 4.

## 8. Conclusions and Future Work

Masking ANN input values for known environmental scenarios has improved the correlation between prediction and outcome by 1.3%. Further work needs to be done to verify that this improvement is consistent in other environments and with masks tuned to other environmental scenarios. This process of masking time regions by groupings of high and low prediction error identifies the dawn and dusk regions of time for a location. Future research could include a model that reevaluates the times for each region and updates the input masks daily; such a model could improve the accuracy of the dawn and dusk time region as changes in season and obstructions impact dawn and dusk.

## Disclosure

Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This work is the direct result of the patience and mentorship of Professor Joel Kubby. The authors would like to thank the University of California, Santa Cruz (UCSC) and the National Aeronautics and Space Administration for the opportunities they have provided. Further thanks are due to Dr. Oscar Azucena, Samuel Kahn, and Steve Willis for their help and expertise. This material is based on work supported by the National Science Foundation under Grant no. 0942439.

## References

[1] Y. Zhang, M. Beaudin, H. Zareipour, and D. Wood, "Forecasting Solar Photovoltaic power production at the aggregated system level," in *Proceedings of the 2014 North American Power Symposium, NAPS '14*, pp. 1–6, 2014.

[2] A. Mellit, M. Benghanem, and M. Bendekhis, "Artificial neural network model for prediction solar radiation data: application for sizing stand-alone photovoltaic power system," in *Proceedings of the IEEE Power Engineering Society General Meeting*, vol. 1, pp. 40–44, June 2005.

[3] H. T. C. Pedro and C. F. M. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," *Solar Energy*, vol. 86, no. 7, pp. 2017–2028, 2012.

[4] E. İzgi, A. Öztopal, B. Yerli, M. K. Kaymak, and A. D. Şahin, "Short-mid-term solar power prediction by using artificial neural networks," *Solar Energy*, vol. 86, no. 2, pp. 725–733, 2012.

[5] Z. Wang and I. Koprinska, "Solar power prediction with data source weighted nearest neighbors," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1411–1418, IEEE, Anchorage, Alaska, Alaska, USA, May 2017.

[6] W. Cabrera, D. Benhaddou, and C. Ordonez, "Solar power prediction for smart community microgrid," in *Proceedings of the 2nd IEEE International Conference on Smart Computing, SMARTCOMP '16*, pp. 1–6, May 2016.

[7] R. Palma-Behnke, C. Benavides, E. Aranda, J. Llanos, and D. Sáez, "Energy management system for a renewable based microgrid with a demand side management mechanism," in *Proceedings of the Symposium Series on Computational Intelligence, IEEE SSCI 2011 - 2011 IEEE Symposium on Computational Intelligence Applications in Smart Grid, CIASG 2011*, pp. 1–8, April 2011.

[8] M. Pourhomayoun, P. Dugan, M. Popescu, and C. Clark, *Bioacoustic signal classification based on continuous region processing, grid masking and artificial neural network*, 2013, https://arxiv.org/abs/1305.3635.

[9] H. Liu, H.-Q. Tian, D.-F. Pan, and Y.-F. Li, "Forecasting models for wind speed using wavelet, wavelet packet, time series and Artificial Neural Networks," *Applied Energy*, vol. 107, pp. 191–208, 2013.