Prediction of hourly solar radiation using temperature and humidity for real-time building energy simulation

Hany Gaballa and Soolyeon Cho*
College of Design, North Carolina State University, Raleigh, NC 27695-7701, USA
*scho3@ncsu.edu

Abstract. Solar radiation is considered one of the most substantial energy sources in our life. This paper discusses how to develop an algorithm to predict real-time hourly solar radiation based on readily available weather data such as temperature and humidity. Artificial Neural Network is one of the most effective technologies for developing algorithms to predict solar radiations. Hidden nodes, learning rates, and epochs are the main three variables. An optimization method is proposed to provide the optimum value of the variables depending on the Coefficient of Variance of the Root Mean Square Error, Normalized Mean Bias Error, and Coefficient of determination.

1. Introduction
Solar radiation plays a significant role in our life as it is utilized in various fields, such as photovoltaics (PV), industry, farming, and building design. If we can have real-time (or even predicted) solar radiation data, it can be better used for daily operations, especially in buildings. In this sense, predicting solar radiation is of major interest to researchers as it is essential to obtain real-time weather data for the prediction of the energy performances in buildings. Temperature and relative humidity can be easily measured in most buildings, while the acquisition of solar radiation data is costly. Also, most weather stations do not have solar radiation measurement equipment.

Researchers developed various approaches to predicting solar radiation. It was found that the prediction accuracy depends on different input parameter combinations. Temperature, humidity, sunshine duration, cloud cover, solar elevation, and air quality; all these parameters used in various algorithms have an effective weight on the output (solar radiation). Through regression modeling analysis done in this paper, it was found that solar radiation could be predicted using time, temperature and relative humidity as inputs which have a significant impact on developing the algorithm. Adding solar zenith angle to the input layer gives more accuracy.

Previous studies show that the Artificial Neural Network (ANN) technology has a powerful ability when dealing with prediction issues [1]. ANN modeling requires input parameters. Through some hidden calculations happened between the input and hidden layer then between the hidden and outputs, it generates output values. This process is called Feed-Forward Neural Network (FFNN). At the end of this process, it gives some weighting factors, which are randomly chosen at the beginning and updated through Back-Propagation (BP) method until the minimum error rate can be found. It was hard to find the optimum parameter values affecting the ANN model manually, so in this study an algorithm was developed with three different variables, giving a range and a distance for each variable to test the whole possible combinations between the three variables automatically. Coefficient of Variance of the Root
Mean Square Error CV(RMSE), Normalized Mean Bias Error (NMBE), and Coefficient of determination (R-squared) was used to measure the accuracy of the model in predicting solar radiations.

The results from this kind of algorithms are affected by the data fed in the algorithm, so in the next sections, all data preparation and simulation process are discussed in detail.

2. Artificial Neural Network
A variety of methods for solar radiation predictions had been used, where ANN is one of the reliable and preferable methods. The results from Alam et al. [2] show that the ANN model gives a good accuracy for hourly diffuse radiation prediction compared to measured data, and better accuracy for monthly mean daily diffuse radiation compared to empirical models. Yadav et al. [1] show that ANN techniques predict solar radiation more accurately in comparison to conventional, linear, nonlinear and fuzzy logic models. According to Qazi et al. [3], ANN is one of the most commonly used methods in predictive data mining. It has evolved as advanced data mining tools in cases where other techniques may not be able to produce acceptable predictive results. ANN provides good accuracy in terms of the Mean Absolute Percentage Error (MAPE) of less than 20%. Voyant et al. [4] mentioned that neural networks had been studied on many parts of the world, and researchers have shown the ability of these techniques to predict the time series of meteorological data accurately. In conclusion, using ANN for prediction requires less formal statistical training, it can solve complex and nonlinear relationships between variables, dependent and independent, and finally can identify all possible interactions between predictor variables [5].

For a better understanding of the ANN model, there is a need to define human brain actions when dealing with problems. Our brain contains billions of neurons, which can process information through electric signals. The neuron’s dendrites receive external data, prepared in its cell body, converted to output, then travel to the next neuron by passing through the Axon. Now, the next neuron can choose to either accept or reject it depending on the strength of the signal [6]. ANN has the same basis as it uses the processing of the brain to develop algorithms to be used for modeling intricate patterns and prediction problems.

ANN model consists of three main stages; input, hidden layers, and output. The input layer is like a bucket that receives input data and passes it to the next layer (hidden layer). The number of input layers is equal to the number of input parameters. The hidden layer is the link between the input and output layers. It contains some neurons which are considered the most crucial part of the ANN model as they imitate the human brain neurons. They have weighted input signals, and through an activation function, they produce an output signal. The output layer contains the results (the predicted values).

Feed-Forward Neural Network is the process of passing parameters from input through the hidden layer to get the output. Feed-Forward Neural Network can learn from examples like human beings learn from their faults, so they do not need a user-specific problem-solving algorithm [7]. But before passing the input values to the hidden layer, there are some weights which are randomly created in the beginning and updated through the whole process to get more predictive results. This process is called Back-Propagation. Through Back-Propagation, an error is computed at the output and distributed backward throughout the network's layers to update initial weights, which is used to train the ANN model. The number of iterations, which the whole calculations of the model should be repeated until getting the least error and high accuracy in the output, is called the number of epochs.

Weights refer to the strength of a connection between two layers; in this model, we have two different weights, first, between the input and hidden layer, and second, between the hidden and output layer. Weights are often initialized with a small random value in the range between 0 and 1.

3. Weather data description
In this model, weather data files are considered the source of input parameters used to train and test the algorithm. The input data is passed through multiple steps to reach its final state to be ready for simulation. The weather data collection station is situated at a latitude of 35.86°N and longitude of 78.78°W, which is Raleigh, North Carolina. The weather in Raleigh is hot and muggy in summers, while
in winters, the daytime is short and cold. Generally, it is wet and partly cloudy year-round. Over the year, the temperature varies from 0.6°C to 31.7°C and is rarely below -6.6°C or above 35°C.

Hourly data of global solar radiation, temperature, relative humidity, and solar zenith angle have been collected from the Typical Meteorological Year (TMY) data for Raleigh-Durham International Airport weather station (Raleigh-Durham.Intl.AP_TMY3). This paper is looking for hourly solar radiation prediction, so it is not logical to consider the nighttime data. For that, all hourly zero values of solar radiation have been removed, which means that the input data has only daytime data.

The data collection started with 8,760 hours as it reflects the whole year data. After removing all zero values, we have 4,356 hours left for daytime data. The data then is divided into four seasons; Spring, Summer, Fall, and Winter to make it more consistent for each season individually. In this paper, Summer data have been chosen for the ANN model inputs. 1,252 Summer hours were divided into training and testing data, 80% for the training and 20% for the testing data.

3.1. Selection of input parameters

This paper focuses on the use of readily available data in the Building Automation System (BAS). Figure 1 shows the correlation between summer hours of solar radiation and easily measured parameters individually; dry-bulb temperature, relative humidity, solar zenith angle, and dew point temperature. Figure 1 shows that dew point temperature does not have meaningful correlation with solar radiation, which is why the first three plots (dry bulb temperature, relative humidity, and solar zenith angle) were chosen to be the input parameters in the ANN model. Dry-bulb temperature, relative humidity, and solar zenith angle have a considerable correlation on solar radiation prediction. Dry-bulb temperature and relative humidity can easily be measured in buildings. Also, the solar zenith angle is easily reachable.

**Figure 1.** Scattered plot for different parameters and hourly global solar radiation (Summertime)

The correlation coefficient (R) for the combination of the previous selected three parameters was found to be 0.81. The values of the month, day, and hour are also added to the input layer.

In this paper, the ANN model uses the sigmoid function as an activate function. The sigmoid function requires a data range from 0 to 1. If a larger input value entered, the ANN becomes saturating, which does not work correctly [8]. So, all the data then were scaled between 0 and 1.

4. ANN model development

Python 3.6 has been used to develop the ANN model. Python is one of the most widely used programming languages for machine learning [9]. ANN model can run complex, nonlinear tasks, and prediction problems. To make it applicable through Python, NumPy and SciPy libraries have been used to construct the ANN model in Python. More information about these libraries can be gathered from Seo et al. [10].
As mentioned before, the ANN model is structured from three layers; the first is the input layer. The input layer consists of several parameters, which in this paper, six parameters; dry-bulb temperature, relative humidity, solar zenith angle, month, day, and hour. In this study, one hidden layer is used to predict hourly global solar radiation, which is the output layer, as shown in Figure 2.

Through this process, there are three variables. First, the number of hidden neurons exist in the hidden layer. From the literature review, it was found that the number of hidden neurons in the ANN hidden layer are changed one by one, and then error rates are evaluated, which is time-consuming [1]. In this study, an optimization technique has been developed for calculating the optimum value of three different variables affecting the ANN model; one of them is the number of hidden neurons.

The second variable is the learning rate, which is considered an essential parameter in the BP process as it is used to settle the changes in the weights at the end of each epoch [11]. The third variable is the number of epochs itself.

![Figure 2. ANN architecture to predict hourly global solar radiation](image)

Three different statistical terms were used to show the prediction accuracy between the measured and predicted values of the ANN model; CV(RMSE), NMBE, and R-squared. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Guideline 14 [12] uses three indices to represent how well a mathematical model describes the variability in measured data. So, in this guideline, it recommends the NMBE value of 10% or less and CV(RMSE) value of 30% or less, that is according to hourly calibration data. For the R-squared value, which indicates correlation, it recommends being 0.80 or higher.

Because of the enormous amount of possible solutions from the three variables combinations, two separate ANN models have been developed. The first ANN model (ANN<sub>1</sub>) is to calculate the optimum values of the variables affecting the ANN model [10]. After each trial, an error rate should be calculated to find the minimum value, and of course, not all possible solutions will be taken into consideration, so, there will be some solutions missing.

Because of the enormous amount of possible solutions from the three variables combinations, two separate ANN models have been developed. The first ANN model (ANN<sub>1</sub>) is to calculate the optimum values of the variables affecting the ANN model [10]. After each trial, an error rate should be calculated to find the minimum value, and of course, not all possible solutions will be taken into consideration, so, there will be some solutions missing.

Because of the enormous amount of possible solutions from the three variables combinations, two separate ANN models have been developed. The first ANN model (ANN<sub>1</sub>) is to calculate the optimum values of the variables affecting the ANN model [10]. After each trial, an error rate should be calculated to find the minimum value, and of course, not all possible solutions will be taken into consideration, so, there will be some solutions missing.

Because of the enormous amount of possible solutions from the three variables combinations, two separate ANN models have been developed. The first ANN model (ANN<sub>1</sub>) is to calculate the optimum values of the variables affecting the ANN model [10]. After each trial, an error rate should be calculated to find the minimum value, and of course, not all possible solutions will be taken into consideration, so, there will be some solutions missing.

Because of the enormous amount of possible solutions from the three variables combinations, two separate ANN models have been developed. The first ANN model (ANN<sub>1</sub>) is to calculate the optimum values of the variables affecting the ANN model [10]. After each trial, an error rate should be calculated to find the minimum value, and of course, not all possible solutions will be taken into consideration, so, there will be some solutions missing.

Because of the enormous amount of possible solutions from the three variables combinations, two separate ANN models have been developed. The first ANN model (ANN<sub>1</sub>) is to calculate the optimum values of the variables affecting the ANN model [10]. After each trial, an error rate should be calculated to find the minimum value, and of course, not all possible solutions will be taken into consideration, so, there will be some solutions missing.

4.1. ANN Optimization process

In most papers, the trial and error process is the most common approach used to find the optimum values of the variables affecting the ANN model [10]. After each trial, an error rate should be calculated to find the minimum value, and of course, not all possible solutions will be taken into consideration, so, there will be some solutions missing.

In this paper, the optimization process took place through ANN<sub>1</sub> to find the optimum variables automatically: the number of hidden neurons, learning rate, and the number of epochs. Each variable of the three was given a range and a step distance. The range of the number of hidden neurons is from 10 to 100 with a distance of 10. The range of learning rate is from 0.01 to 1.00 with a distance of 0.01. The range of the number of epochs is from 100 to 1000 with a distance of 100.
All possible solutions between the three variables have been simulated. The three statistical methods for assessing the performance of the prediction model also were calculated for each step, CV(RMSE), NMBE, R-squared. All values of the three variables and the three statistical methods were printed out in an excel file. The minimum CV(RMSE) was selected, looking at NMBE and R-squared to make sure that the values within the range of ASHRAE Guideline 14.

Now, it is time for ANN₂ to use these values to find the most accurate solar radiation prediction with minimum error. It is possible to combine the two ANN models into one, but it was preferred to separate the two models to reduce the time consumed in simulations as there is no need to have all other predicted values of solar radiation with high error rates, just to focus on the optimum one. So, both ANN₁ and ANN₂ have the same inputs and output data as shown in Figure 2. The main differences are, first the calculation process if it is a loop between the three variables (as in ANN₁) or constant (as in ANN₂), and second the printout values if it is the three variables (as in ANN₁) or just solar radiation (as in ANN₂).

5. Results and discussion
Through 10,000 trials simulated, the results from ANN₁ shows that the optimum values are 40, 0.84, and 900 for the number of hidden neurons, number of epochs, and learning rate respectively with CV(RMSE) of 19.06%, NMBE of 5.43%, and R-squared value of 0.89. In reference to ASHRAE Guideline 14, the results meet the tolerance requirements and lie within the acceptable tolerance ranges.

![Figure 3](image-url)

**Figure 3.** Comparison between the predicted and measured hourly global solar radiation

The optimum values of the three variables are then taken to be constant in the ANN₂ to predict the solar radiation. Figure 3 shows a comparison between the predicted and measured hourly global solar radiation. The horizontal axis indicates the number of tested hours which is in this case 250 hours. These hours represent randomized 20% of the Summer data used in this algorithm. The other 80% is used in the training process. The vertical axis shows the global solar radiation values.

![Figure 4](image-url)

**Figure 4.** Regression analysis plot for the optimum model between output (Predicted) and target of hourly global solar radiation
The scattered plot in Figure 4 shows the correlation between the predicted and measured solar radiation data. The x-axis shows the measured data, while the y-axis for the predicted values. The dashed line shows the consistency between the two values with an R-squared value of 0.89.

At the end of this step, we can say that the optimization process worked well and achieved its goal of predicting solar radiation using the readily available measured parameters inside buildings. It gives the acceptable performance and accurate predictions. The number of hidden neurons, epochs, and learning rates has a good relationship with the performance of the ANN model and that appears in different values of CV(RMSE), NMSE, and R-squared.

6. Conclusion

It was found from the literature review that ANN modeling is one of the effective ways used in predictions. In this study, a two-step methodology for choosing the ANN model to predict hourly global solar radiation was presented. The objective of this paper was to focus on the available meteorological data in the building such as temperature and relative humidity to predict solar radiation. The solar zenith angle has been added to the input parameters, which is easily reachable.

Using the optimization process for training the algorithm (ANN₁) may take quite a long time to find the optimum values of the three variables; however, as it is found once, it can be used in prediction algorithm (ANN₂) easily with minimal time.

The CV(RMSE), NMSE and R-squared value are found to be 19.06%, 5.43%, and 0.89, respectively, which shows a good prediction accuracy.

For future work, one more hidden layer can be added for better prediction. It will also affect the number of hidden neurons in each hidden layer and other variables. It will be more complicated with a four-layer ANN model. There should be a comparison between the three-layer and four-layer ANN models to find the optimum and most accurate algorithm and see the effectiveness of the new model.

References
[1] Yadav A K and Chandel S S 2014 Solar radiation prediction using artificial neural network techniques: a review Renewable and Sustainable energy reviews 33 p 772-81
[2] Alam S, Kaushik S C and Garg S N 2009 Assessment of diffuse solar energy under general sky condition using artificial neural network Applied Energy 86 p 554-64
[3] Qazi A, Fayez H, Wadi A, Raj R G, Rahim N A and Khan W A 2015 The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review Journal of cleaner production 104 p 1-12
[4] Voyant C, Notton G, Kalogirou S, Nivet ML, Paoli C, Motte F and Fouilloy A 2017 Machine learning methods for solar radiation forecasting: A review Renewable Energy 105 p 569-82
[5] Mohanty S, Patra P K and Sahoo S S 2016 Prediction and application of solar radiation with soft computing over traditional and conventional approach–A comprehensive review Renewable and Sustainable Energy Reviews 56 p 778-96
[6] Mahanta J 2017 Introduction to Neural Networks Advantages and Applications Towards Data Science
[7] Mellit A, Eleuch H, Benghanem M, Elaoun C and Pavan A M 2010 An adaptive model for predicting of global, direct and diffuse hourly solar irradiance Energy Conversion and Management 51 p 771-82
[8] Rashid T 2016 Make your own neural network CreateSpace Independent Publishing Platform
[9] Guo P 2014 Python is now the most popular introductory teaching language at top us universities BLOG@ CACM July 47
[10] Seo B, Yoon Y, Song S and Cho S 2019 ANN-based thermal load prediction approach for advanced controls in building energy systems ARCC International Conf.
[11] Rezrazi A, Hanini S and Laidi M 2016 An optimisation methodology of artificial neural network models for predicting solar radiation: a case study Theoretical and applied climatology 123 p 769-83
[12] A Guideline 14 2014 Measurement of Energy, Demand and Water Savings American Society of Heating Ventilating and Air Conditioning Engineers

6