Theory-Guided AI for Intraday Solar Radiation Prediction

Completed Research Paper

Hashai Papneja
University of Georgia
hashai@uga.edu

Abstract

The importance of incorporating underlying theory and domain knowledge while building artificial intelligence-based predictive models is examined. Using the context of predicting intraday solar radiation, we show that a theoretically grounded predictive model yields better performance and offers more interpretability and generalizability than a model that relies solely on other variables. Inclusion of theoretically-guided variables in data-driven predictive models is proposed as a means to mitigate overfitting and reduce potential bias.

Keywords

Renewable energy, solar radiation prediction, neural networks, theory, algorithmic bias.

Introduction

Accurate prediction of intraday solar radiation lies at the heart of several renewable energy initiatives. Forecasted solar radiation is an important parameter, for instance, in several photovoltaic (PV) power generation systems where forecasts are required to mitigate problems such as voltage fluctuations and grid losses that arise with increased PV integration. PV-generated solar power has seen tremendous growth over the last decade, with a total of 402.5 GW installed globally at the end of 2017—a 33% increase since 2016 alone (PVPS 2018), and minimizing integration-related costs has become ever more important. Besides PV applications, an increasing number of artificial lighting control systems, particularly in the context of controlled environment agriculture, also seek to accurately predict incident solar radiation in order to optimize electricity consumption and minimize costs (Albright et al. 2000; Watson et al. 2018). Several control systems in intelligent buildings also involve forecasting temperature and solar irradiance (Argiriou et al. 2004).

Literature on intraday prediction of solar radiation ranges from earlier applications of statistical modeling techniques such as hidden Markov chains and autoregressive time series forecasts, to more recent applications of artificial intelligence (AI) techniques such as artificial neural networks (ANNs) and support vector machines (SVMs). While such techniques have grown in popularity over the last decade due to an abundance of both data and computational power, parsimony and interpretability of these "black-box" models remains a pressing issue (Lipton 2018). Perils of deviations from model parsimony and interpretability include learning spurious relationships, drawing misleading conclusions, a lack of generalizability, transparency, and outcome explainability. These are especially important, for instance, when dealing with critical, high risk problems such as those in the healthcare industry, or when operating in regulated industries such as financial services. Well-known examples of negative outcomes or unintended algorithmic biases arising from black-box models include the case of Amazon's AI hiring model that biased decisions against women, and the theory-agnostic Google Flu Trends model (Lazer et al. 2014).

Predictive analytics can add theoretical and practical value to IS research (Shmueli and Koppius 2011). Yet, IS scholars should be vigilant to issues such as selection biases and generalizability that can pose a threat to valid causal inferencing when using data-driven machine learning techniques (Rai 2016). Further, predictive models should be understandable to decision makers (Baesens et al. 2016). This study explores the importance of incorporating underlying theory in the predictive model-building process. Specifically, it investigates the use of neural networks in predicting intraday solar radiation. However, unlike existing

literature on the subject which heavily focuses on various meteorological variables, weather forecasts, and satellite imagery, this study relies primarily on the underlying theoretical mechanism—the theoretical extraterrestrial solar radiation curve at the given location—to make predictions. In doing so, it attempts to build a more parsimonious and interpretable model. Thus, the research question that this study aims to answer is as follows:

RQ: How can intraday solar radiation be predicted using a theory-guided, parsimonious model?

Literature Review

Early work on solar radiation forecasting relied on statistical modeling techniques. Aguiar et al. (1988) employ Markov transition matrices to generate sequences of daily global radiation values for any location, using as input only the average monthly radiation for that location. The method is based on the observation that there is a significant correlation only between radiation values for consecutive days, and that the probability of occurrence of radiation values is the same for months with the same clearness index—the ratio of surface solar radiation to extraterrestrial solar radiation. Graham and Hollands (1990) employ a stochastic disaggregation procedure to generate synthetic sets of hourly solar irradiation values using only the 12 monthly means of daily atmospheric transmittance (akin to clearness index) values. Analyses of meteorological records revealed that hourly solar radiation can be closely predicted using the atmospheric transmittance for the day. Aguiar and Collares-Pereira (1992) present a Time-dependent Autoregressive Gaussian (TAG) model for the generation of synthetic sequences of hourly horizontal global radiation. The model takes into account the fact that the daily clearness index sequences change not only with day, but also with the solar hour. Skartveit and Olseth (1992) model the probability distribution and lag-1 autocorrelation of short term irradiance data (1–10 minutes) and use these in a first order autoregressive model for the synthetic generation of short term data.

Most of the early studies covered above make use of the clearness index values (or atmospheric transmittance—a similar concept) to predict solar radiation at a given location. Clearness index is the ratio of global horizontal irradiance (GHI) to extraterrestrial solar radiation (ETR) on a horizontal surface—concepts explained in the next section. This variable is theoretically informed in that it incorporates information from the extraterrestrial solar radiation curve of that particular location. However, as the following paragraphs elaborate, most of the later studies relying on AI techniques seem to have diverted the focus away from theoretically informed variables to a variety of other, theory-agnostic meteorological variables and to computationally sophisticated techniques such as predicting cloud motion vectors through ground and satellite imagery. In a way, this study attempts to restore focus to underlying theoretical mechanisms and variables by demonstrating their potential to significantly increase predictive accuracy while also improving model parsimony, interpretability, and generalizability.

Among recent studies on solar radiation forecasting, Mellit et al. (2006) investigate the use of feed-forward artificial neural networks using wavelets as activation functions to predict daily total solar radiation. Previous values of daily total solar radiation data are used as inputs to predict the same for subsequent days. Data recorded from 1981 to 2001 by a meteorological station in Algeria are used. The model predicts daily total solar radiation values with a mean absolute percentage error (MAPE) of 6% or less. Reikard (2009) evaluates the ability of several types of time series models to predict radiation at ground level. The study finds that the choice of the model depends on the forecast resolution. At lower resolutions (hours), the data is dominated by the diurnal cycle, where autoregressive integrated moving average (ARIMA) models do better. At higher resolutions (minutes), the data is more dominated by short-term patterns which can be picked up by regressions on levels or artificial neural networks. The study uses cloud cover, humidity, atmospheric turbulence, and lagged GHI values as inputs to develop neural network models that predict hourly solar radiation. The models yield MAPE values ranging from 22.14% to 63.75% for 30-minute ahead predictions. Paoli et al. (2010) develop a methodology for predicting daily global solar radiation on a horizontal surface. The study proposes a pre-processing approach based on extraterrestrial normalization and finds that this reduces NRMSE by about 6% (NRMSE = 15% for summer months, 37% for winter) compared to conventional prediction methods of ARIMA, Markov chains, Bayesian inference, and *k*-Nearest Neighbors. Chen et al. (2013) present an intraday solar radiation forecast technique based on fuzzy logic and neural networks. By using fuzzy logic to classify the day as sunny, cloudy, or rainy, and then using the neural network to predict solar radiation, the study yields MAPE values ranging from about 6% to 9.65% under different sky and temperature conditions. However, the study does not utilize the theoretical

extraterrestrial solar radiation values as an input to the model, relying heavily on weather forecasts and temperature. Yadav et al. (2014) review develop three neural network models using the Waikato Environment for Knowledge Analysis (WEKA) software based on the following input parameters—latitude, longitude, temperature, maximum temperature, minimum temperature, altitude and sunshine hours. The maximum MAPE for the three models are found to be 20.12%, 6.89%, and 9.04%. WEKA identifies temperature, maximum temperature, minimum temperature, altitude and sunshine hours as the most relevant input variables and latitude, longitude as the least influencing variables. Wollsen and Jørgensen (2015) propose a method for predicting short-term (less than 25 hours ahead) solar irradiance and outdoor temperature using a nonlinear autoregressive with external input (NARX) artificial neural network that only requires previous measurements of the parameters (solar radiation and whether it is day or night) as input. The model is found to outperform a commercial forecast, yielding an RMSE of 85.8 W/m^2 for 1-step ahead (hourly) predictions. We come across only one study that makes use of the extraterrestrial solar radiation values while making intraday predictions—Akarslan and Hocaoglu (2016). This study employs different predictive models for different seasons of the year, yielding NRMSE values that range from about 34.87% to 40.45% for the different locations tested.

In addition to the above studies, two articles that review extant literature on machine learning (ML) techniques used in solar radiation forecasting are worth noting. Lauret et al. (2015) compare supervised ML techniques—models where the outcome variable is known—and find that NRMSE values for neural network models range from 19.65% to 25.99%. Similarly, Voyant et al. (2017) find that NRMSE values for neural network models range from 18% to 24%.

The use of different measures of predictive accuracy, different forecast horizons, and different time-step resolutions across the studies makes it difficult to establish a common benchmark against which to compare model performance. Table 1 below summarizes the above-mentioned recent studies that predict solar radiation, salient characteristics of the models employed, and the reported predictive accuracy.

Study	Prediction Horizon	Model Characteristics	Reported Predictive Accuracy
Mellit et al. (2006)	Daily total solar radiation	Wavelet neural network (NN) models that use first *n* values of total solar radiation from preceding months.	MAPE ≤ 6%
Reikard (2009)	Intraday solar radiation from 5 min to 4 hours ahead.	ARIMA and NN models that use cloud cover, humidity, atmospheric turbulence, and previous GHI values.	NN better than ARIMA at higher resolutions. MAPE from 22.14% to 63.75% for 30 min ahead predictions.
Paoli et al. (2010)	Daily total solar radiation	Preprocessed data fed to NN models that use clearness index, previous GHI values, ETR normalization, and a seasonal factor value for each day.	NN better than ARIMA, Bayesian inference, Markov chains, and *k*-NN. NRMSE from 15% to 37%.
Chen et al. (2013)	Intraday hourly solar radiation	Uses fuzzy logic to determine sky conditions and then feeds as input to NNs previous solar radiation values, current and forecast sky conditions, and current and forecast temperature.	MAPE from 6% to 9.65%
Yadav et al. (2014)	Monthly average solar radiation	NN models that use latitude, longitude, temperature, maximum temperature, minimum	MAPE from 6.89% to 20.12%

		temperature, altitude and sunshine hours.	
Wollsen and Jørgensen (2015)	Intraday hourly solar radiation	NN model that uses previous solar radiation values, and a binary input signifying whether it is day or night.	RMSE = 85.8 W/m^2
Akarslan and Hocaoglu (2016)	Intraday hourly solar radiation	A hybrid approach tested across 3 locations that uses the clearness index to decide whether to predict solar radiation using a linear combination of previous values, or as a combination of previous and ETR values.	NRMSE from 34.87% to 40.45%
Literature Reviews / Meta-analyses:			
Lauret et al. (2015)	Various	Various machine learning (ML) based models	NRMSE from 19.65% to 26%
Voyant et al. (2017)	Various	Various ML-based models	NRMSE from 18% to 24%

Table 1. Summary of Recent Studies on Solar Radiation Prediction

Whereas most previous studies use a combination of weather variables and forecasts to predict intraday solar radiation, the purpose of this study is to reveal the importance of the underlying theoretical mechanism which forms the basis of variation in intraday solar radiation. Incident solar radiation is a function of extraterrestrial solar radiation, a quantity that varies predictably as the distance between the location on Earth and the Sun varies throughout the day, and throughout the year. A predictive model that incorporates this information is expected to yield better accuracy than a model that relies solely on other factors. This is our primary proposition. Such a model can also help improve interpretability of the solution, and improve generalizability in that extraterrestrial solar radiation can be calculated accurately for any geographical location without the need for sophisticated instruments.

The remainder of this paper is organized as follows. Relevant concepts relating to extraterrestrial solar radiation are covered in the next section. A description of the methodology is then provided wherein details about data collection, cleansing, and model implementations are described. Results from each of the models are then presented, followed by their interpretation and discussion. Limitations and directions for future research are discussed, and we conclude with key implications.

Extraterrestrial Solar Radiation

Extraterrestrial solar radiation (ETR), also known as top-of-atmosphere irradiance, refers to the amount of sunlight incident at a location if there were no atmosphere or clouds present[1]. It varies as a function of latitude of the location and time of year. Figure 1 below shows ETR for Athens, Georgia, USA for two different days of the year—Day 1 and Day 180. The number of observations is constant across the two graphs. The graphs show that day length is shorter on Day 1 (winter in Athens, GA) than on Day 180 (summer), and that peak sunlight received is higher in summers. The curves would vary for locations on a different latitude.

[1] Source: NREL Solar Resource Glossary available at https://www.nrel.gov/grid/solar-resource/solar-glossary.html; accessed February 28, 2019.

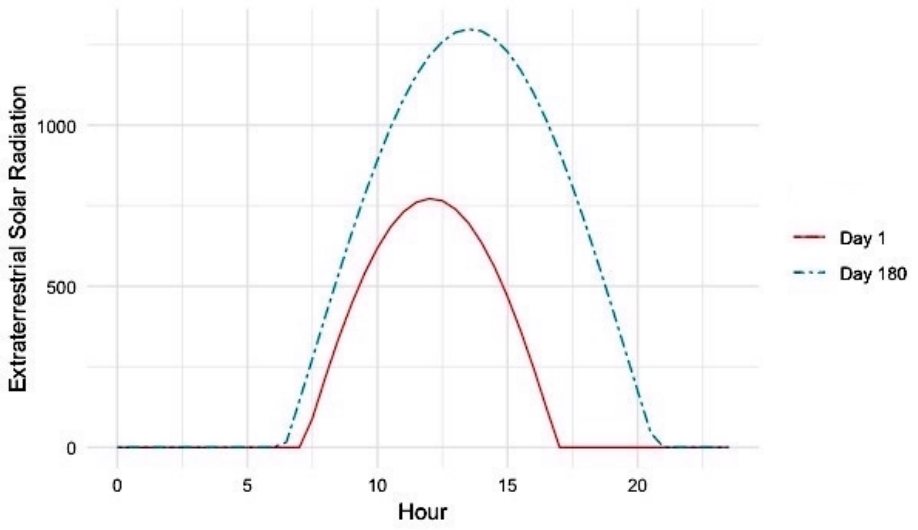

Figure 1: Extraterrestrial Solar Radiation Values for Athens, GA for Day 1 and Day 180

Various atmospheric effects attenuate ETR as it reaches ground. Among these are cloud cover, moisture content, and other atmospheric contents such as aerosols and particulate pollution. As a result, some sunlight gets absorbed and diffused in the atmosphere before reaching ground. The amount of sunlight measured at the surface of the Earth at a given location with a surface perpendicular to the Sun is called direct normal irradiance (DNI), or beam radiation. It is equal to the ETR minus any atmospheric losses due to absorption and diffusion. Diffuse horizontal irradiance (DHI) is the radiation incident on the Earth's surface from sunlight scattered by the atmosphere. Global horizontal irradiance (GHI) is the total irradiance (includes DNI and DHI) from the Sun on a horizontal surface on Earth and is the variable of interest of this study. GHI is measured by a pyranometer mounted horizontally on the surface of the Earth.

Methodology

Since the subject of our study is a continuous variable—incident solar radiation—the problem lends itself well to a supervised learning approach, specifically to artificial neural networks (ANNs). This study focuses on a widely-used type of ANN, the feedforward neural network (FNN). FNNs are networks wherein the output of the network is a function of only the current input, and not any past input or past outputs. In contrast, the output of a recurrent neural network (RNN) is a function of both the current input and the previous state of the (hidden layer of the) network. Neural networks are a non-parametric method that make no statistical assumptions about the data and are able to deal with non-linear problems. They are widely known to tackle complex, ill-defined problems once trained to predict outcomes from examples.

The primary proposition guiding our research is that a theory-driven model will yield better predictions and offer better interpretability and generalizability than a naïve model that relies solely on other variables. To probe this, four models are constructed: (i) a naïve model that only consists of readily-available meteorological variables as predictors, (ii) a theoretical-only model that only consists of ETR as the predictor, (iii) a combined model that uses both, naïve variables and the theoretically-guided variable, and (iv) a parsimonious combined model that attempts to reduce the number of naïve variables while still maintaining predictive accuracy. An intraday time interval of 30 minutes was used. Each model attempts to predict incident solar radiation one time-step ahead, i.e., 30 minutes into the future.

Data Collection

Weather and intraday solar radiation data for Athens, GA were obtained from the University of Georgia's Climatology Lab. Four consecutive years of data—2015 to 2018—were used for the purposes of this study. Observations were gathered in 30-minute time intervals starting at 00:00 hours to 23:30 hours for each day. Each observation consists of temperature (˚F), relative humidity (%), wind speed (mph), raw

barometric pressure (mb), and solar radiation (W/m²) values at that particular point in time. Intraday ETR values (W/m²) for Athens, GA for each of the four years were calculated using the *solaR* package (Perpiñán 2012) in *R*. These were corrected for Daylight Savings Time and then matched with the four years of 30-minute observations from the Climatology Lab. Actual solar radiation readings of more than the theoretical extraterrestrial solar radiation value were corrected by setting them to the theoretical maximum. There were 457 such readings (0.66% of total), with an average magnitude of error of about 7.5 W/m² (4.37% of mean value). After accounting for missing data, the final dataset consisted of 68,936 observations. Ideally, the number of observations would be 70,128 (48 readings per day for 1,461 days), implying a 1.7% loss.

Model Implementations

While neural networks are well-known for their ability to "learn" complex relations, they may sometimes overfit the training data or converge to a local optimum, yielding unstable or potentially biased predictions (Geva et al. 2017). One way to mitigate overfitting is to test the model on out-of-sample data. Hence the original dataset was partitioned into training and testing datasets whereby the first three years (2015 to 2017) were used for training and cross-validation, and the fourth year (2018) was used for testing. The training dataset was shuffled for efficient convergence (LeCun et al. 1998). Values of each of the model inputs were standardized. Since test values are not to be known beforehand, input values of the test dataset were standardized using the mean and standard deviation of the training dataset.

A 4-fold cross-validation (using 25% of the training dataset) was implemented to reliably evaluate each model and identify the early stopping point while training the model. Cross-validation consists of splitting the training dataset into *K* partitions, training the model on *K-1* partitions while evaluating it on the remaining partition (Chollet and Allaire 2018). The averages of the *K* validation scores (RMSE, NRMSE, and MAE—see Table 2 below) are then reported.

Repeated fine-tuning of the neural network parameters (specifically, weights) over multiple epochs can lead to overfitting the network to the training data. Stopping the training process after a suitable number of epochs helps prevent overfitting (Chollet and Allaire 2018). Hence, an early stopping criteria that minimizes the average Mean Absolute Error (MAE) (across previous 10 epochs) was implemented.

The *keras* package[2] in *R* with TensorFlow as the backend was used to build the neural network models. Mean Squared Error (MSE) was used as the loss function. *keras* uses a modified version of the backpropagation algorithm, RMSProp, as the optimizer, with a default learning rate of 0.001. A densely connected network comprising of 2 hidden layers with 16 nodes each was used across all models. The Rectified Linear Unit (RELU) activation function was used due to its sparse activation characteristic. Sparse activation results in concise models that are less prone to overfitting (Chollet and Allaire 2018).

It is pertinent to note that, in accordance with the primary goal of this study, model structure (the number of hidden layers, nodes per layer, connections, and activation functions), parameter values, and early stopping criteria were held constant across all models so that differences in model output are a result of modifying only the model inputs. This preserves model comparability. In an effort to foster comparability across studies, three performance metrics are evaluated and presented for each model. Table 2 below shows the three performance metrics and their corresponding formulae.

Term	Meaning	Formula
RMSE	Root Mean Squared Error, expressed in units of the predicted variable	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}}$
NRMSE	Normalized RMSE, expressed as a percentage	$NRMSE = \frac{RMSE}{\bar{y}}$

[2] Allaire, J.J., and Chollet, F. 2018. "keras: R Interface to 'Keras'" available at https://CRAN.R-project.org/package=keras; accessed August 2, 2018.

MAE	Mean Absolute Error, expressed in units of the predicted variable	$MAE = \frac{\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert}{n}$

Table 2. Performance Metrics Used

Results

The subsections below describe the results obtained from running the above-mentioned models. In an effort to rule out severe overfitting, results for both the training (in-sample) and testing (out-of-sample) datasets are evaluated and presented below.

Model 1: Naïve Model

The least theoretically-guided model—the naïve model—takes only readily-available weather variables as input. These include wind speed, relative humidity, temperature, and barometric pressure. Table 3 below shows results for the naïve model.

Metric	Training	Testing
RMSE (W/m²)	202.63	202.85
NRMSE	57.65%	60.60%
MAE (W/m²)	143.34	137.39

Table 3. Predictive Accuracy: Model 1 (Naïve)

Model 2: Theoretical-Only Model

The theoretical-only model relies on a single theoretically-guided variable—the ETR value—as input. Though it is unconventional to build a neural network with only one input, evaluation of such a model helps probe the efficacy of using underlying theory as the sole driver of predictive accuracy. Table 4 below shows results for the theoretical-only model.

Metric	Training	Testing
RMSE (W/m²)	133.47	134.46
NRMSE	37.97%	40.16%
MAE (W/m²)	72.88	73.62

Table 4. Predictive Accuracy: Model 2 (Theoretical-only)

As seen from Table 4 above, a theoretical-only model yields better predictive accuracy than a naïve model. RMSE decreases from about 202 W/m² for the naïve model to about 134 W/m² for the theoretical-only model—a 33% decrease. This offers support for our proposition that incorporating the underlying theoretical mechanism in predictive model-building will significantly improve predictive accuracy.

Model 3: Combined Model

The combined model combines all the naïve input variables of Model 1 and the theoretically-guided variable of Model 2 to probe the efficacy of a combined approach to predictive modeling. Table 5 below shows results for the combined model.

Metric	Training	Testing
RMSE (W/m²)	103.49	106.11
NRMSE	29.44%	31.70%
MAE (W/m²)	50.86	53.52

Table 5. Predictive Accuracy: Model 3 (Combined)

As seen from Table 5 above, the combined model results in a further 21% reduction in RMSE as compared to the theoretical-only model (Table 4), and about a 48% reduction in RMSE as compared to the naïve model (Table 3). This offers further support for our proposition.

Model 4: Parsimonious Combined Model

The parsimonious combined model aims to offer better parsimony and hence better interpretability than the combined model above. It aims to minimize the number of inputs (the feature set) while still attaining a predictive accuracy close to the combined model. To this end, literature on solar radiation is examined in an attempt to understand which input variables should be excluded. Accordingly, both wind speed and barometric pressure are excluded as they offer lesser predictive ability than humidity and temperature. Model 4 thus comprises of two naïve variables—humidity and temperature—along with the theoretical variable—ETR. Table 6 below shows results for this parsimonious combined model.

Metric	Training	Testing
RMSE (W/m²)	103.40	105.25
NRMSE	29.42%	31.45%
MAE (W/m²)	50.92	50.17

Table 6. Predictive Accuracy: Model 4 (Parsimonious Combined)

As seen from Table 6 above, predictive accuracy of the parsimonious combined model comprising of three input variables—ETR, relative humidity, and temperature—is about the same as that for the combined model comprising of five input variables.

Discussion

The naïve model consisting of only theoretically agnostic weather variables and yields an NRMSE of 60.6% on the test dataset. In contrast, the theoretical-only model yields an NRMSE of about 40.16% — a 33% reduction in prediction error. The combined model further reduces prediction error by 21% to yield an NRMSE of 31.7%. However, a parsimonious model which incorporates two out of the four naïve variables—temperature and humidity—along with the theoretical variable—ETR yields comparable predictive accuracy (NRMSE of 31.45%) to the combined model. Thus, the study finds support for the proposition that theoretically guided predictive models yield better predictive accuracy, and can offer parsimony and hence better interpretability than models that relies solely on naïve, theory-agnostic variables. The inclusion of theoretically guided variables in data-driven predictive models can help mitigate algorithmic biases.

To the extent that costs matter, the ETR values offer a simple and cost-effective way to improve predictive accuracy of intraday solar radiation prediction models. Improved predictive accuracy reduces supply-planning and grid-integration costs for renewable energy providers. Further, ETR values can be calculated

easily for any location on the planet without the use of expensive instruments. For a geographically dispersed provider, this affords generalizability, and is also pertinent when ongoing costs of maintaining the instruments are taken into consideration—it can directly impact the provider's bottom line.

Limitations and Future Directions

There are several limitations to this study. First, we make use of only readily-available meteorological variables in order to construct the naïve model. This feature set can be considered insufficient, and this could be the reason behind the poor predictive accuracy of the naïve model. This is a valid concern, given that we do not input a measure for 'time of day'—a widely used variable in predicting intraday solar radiation—to the naïve model. We run an additional model that incorporates this measure as an additional input and find that predictive accuracy does increase, but still fails to surpass the level offered by the theoretically-informed parsimonious combined model (Model 4). Table 7 below shows performance metrics for this model (Model 5) which includes the following five variables as input—wind speed, relative humidity, temperature, barometric pressure, and elapsed minutes since midnight. One can see that the predictive accuracy of the final model above—Model 4—is better than that of Model 5. Specifically, Model 4 yielded RMSE, NRMSE, and MAE values of 105.25 W/m², 31.45%, and 50.17 W/m² respectively on the test dataset.

Metric	Training	Testing
RMSE (W/m²)	114.98	112.59
NRMSE	32.71%	33.64%
MAE (W/m²)	62.36	58.02

Table 7. Predictive Accuracy: Model 5

Another possible limitation could be that the study only predicts solar radiation for one time-step ahead, i.e., 30 minutes into the future. In contrast, some other studies have predicted solar radiation on an hourly or a daily basis. However, we do not expect the primary finding to change if this were the case. We expect the theoretically-driven model to outperform the naïve model even when predicting solar radiation for larger timeframes. Further research in this direction could corroborate our finding.

Lastly, our study was limited to the context of predicting intraday solar radiation. Further research across other predictive modeling domains would help bring to light the external validity of our finding.

Conclusion

Theoretical reasoning has become an often-overlooked part of the predictive equation, given the abundance of data and computational power. This study explores the use of underlying theory in building predictive models using artificial neural networks. It attempts to build a theory-guided, interpretable and parsimonious model to predict intraday solar radiation. Various models are explored, and among the models explored, the theoretically-informed model that incorporates previous (one time-lagged) solar radiation yields the best predictive accuracy. Implications for renewable energy initiatives that utilize solar radiation predictions were briefly discussed.

This study also underscores the importance of pertinent domain knowledge when building predictive models. Such knowledge, when coupled with big data and machine learning techniques can vastly improve predictive outcomes, lend much-needed interpretability to the solution, and improve the generalizability of the solution.

REFERENCES

Aguiar, R. J., and Collares-Pereira, M. 1992. "TAG: A Time-Dependent, Autoregressive, Gaussian Model for Generating Synthetic Hourly Radiation," *Solar Energy* (49:3), pp. 167-174.

Aguiar, R. J., Collares-Pereira, M., and Conde, J. P. 1988. "Simple Procedure for Generating Sequences of Daily Radiation Values Using a Library of Markov Transition Matrices," *Solar Energy* (40:3), pp. 269-279.

Akarslan, E., and Hocaoglu, F. O. 2016. "A Novel Adaptive Approach for Hourly Solar Radiation Forecasting," *Renewable Energy* (87), pp. 628-633.

Albright, L. D., Both, A. J., and Chiu, A. J. 2000. "Controlling Greenhouse Light to a Consistent Daily Integral," *Transactions of the ASAE*, pp. 421-431.

Argiriou, A. A., Bellas-Velidis, I., Kummert, M., and André, P. 2004. "A Neural Network Controller for Hydronic Heating Systems of Solar Buildings," *Neural Networks* (17:3), pp. 427-440.

Baesens, B., Bapna, R., Marsden, J.R., Vanthienen, J., and Zhao, J.L. 2016. "Transformational Issues of Big Data And Analytics in Networked Business," *MIS Quarterly* (40:4), pp. 807-818.

Chen, S. X., Gooi, H. B., and Wang, M. Q. 2013. "Solar Radiation Forecast Based on Fuzzy Logic and Neural Networks," *Renewable Energy* (60), pp. 195-201.

Chollet, F., and Allaire, J.J. 2018. *Deep Learning With R*, Shelter Island, NY: Manning Publications.

Geva, T., Oestreicher-Singer, G., Efron, N., and Shimshoni, Y. 2017. "Using Forum and Search Data for Sales Prediction of High-Involvement Projects," *MIS Quarterly*, (41:1), pp. 65-82.

Graham, V. A., and Hollands, K. G. T. 1990. "A Method to Generate Synthetic Hourly Solar Radiation Globally," *Solar Energy* (44:6), pp. 333-341.

Lauret, P., Voyant, C., Soubdhan, T., David, M., and Poggi, P. 2015. "A Benchmarking of Machine Learning Techniques for Solar Radiation Forecasting in an Insular Context," *Solar Energy* (112), pp. 446-457.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. 2014. "The Parable of Google Flu: Traps in Big Data Analysis," *Science* (343), pp. 1203-1205.

LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. 1998. "Efficient BackProp," *Neural Networks: Tricks of the Trade*, London, UK: Springer-Verlag.

Lipton, Z. C. 2018. "The Mythos of Model Interpretability," *Communications of the ACM* (61:10), New York, NY, USA: ACM, pp. 36-43.

Mellit, A., Benghanem, M., and Kalogirou, S.A. 2006. "An Adaptive Wavelet-Network Model for Forecasting Daily Total Solar-Radiation," *Applied Energy*, (83:7), pp. 705-722.

Paoli, C., Voyant, C., Muselli, M., and Nivet, M. L. 2010. "Forecasting of Preprocessed Daily Solar Radiation Time Series Using Neural Networks," *Solar Energy* (84:12), pp. 2146-2160.

Perpiñán, L. O., 2012. "solaR: Solar Radiation and Photovoltaic Systems with R," *Journal of Statistical Software* (50:9), pp. 1-32.

PVPS, I. 2018. "A Snapshot of Global Photovoltaic Markets (1992-2017)," *International Energy Agency (IEA)*, pp. 1-15.

Rai, A., 2016. "Editor's Comments: Synergies Between Big Data and Theory," *MIS Quarterly* (40:2), pp. iii-ix.

Reikard, G. 2009. "Predicting Solar Radiation at High Resolutions: A Comparison of Time Series Forecasts," *Solar Energy* (83:3), pp. 342-349.

Shmueli, G. and Koppius, O.R., 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553-572.

Skartveit, A., and Olseth, J. A. 1992. "The Probability Density and Autocorrelation of Short-Term Global and Beam Irradiance," *Solar Energy* (49:6), pp. 477-487.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., and Fouilloy, A. 2017. "Machine Learning Methods for Solar Radiation Forecasting: A Review," *Renewable Energy*, pp. 569-582.

Watson, R. T., Boudreau, M.-C., and van Iersel, M. W. 2018. "Simulation of Greenhouse Energy Use: An Application of Energy Informatics," *Energy Informatics* (1:1), Energy Informatics, pp. 1-14.

Wollsen, M. G., and Jørgensen, B. N. 2015. "Improved Local Weather Forecasts Using Artificial Neural Networks," in *Advances in Intelligent Systems and Computing* (Vol. 373), pp. 75-86.

Yadav, A. K., Malik, H., and Chandel, S. S. 2014. "Selection of Most Relevant Input Parameters Using WEKA for Artificial Neural Network Based Solar Radiation Prediction Models," *Renewable and Sustainable Energy Reviews*, pp. 509-519.