COBEM 2021 26th International Congress of Mechanical Engineering
ABCM Associação Brasileira de Engenharia e Ciências Mecânicas
November 22-26, 2021, Virtual Congress, Brazil

COB-2021-1090
SOLAR IRRADIANCE FORECASTING USING MACHINE LEARNING, COMPARISON OF XGBOOST AND BOOSTING MODELS

David Mickely Jaramillo Loayza
Paulo Alexandre Costa Rocha
Maria Eugênia Vieira da Silva
Mechanical Engineering Department, Technology Center, Federal University of Ceará, Fortaleza, CE 60020-181, Brazil
mickely.jaramillo@outlook.com
paulo.rocha@ufc.br
eugenia@ufc.br

Abstract. *One of the most important tools for electricity generation based on solar energy is the solar irradiance forecasting. This is because the solar irradiation depends on mostly by weather conditions variability, and an accurate forecasting is needed to control and distribute efficiently the electricity demand of generation systems. Machine learning models are being used to develop many types of forecasting models, namely the Extreme Gradient Boosting (XGBoost) model, which has high capability and works faster than the others tree boosting models. Many forecasting models that are robust have a significant demand of resources to work, hence representing a high cost of implementation. On the other hand, the XGBoost is a model that has high efficiency and low resources demand. That capability allows to test models with different time scales, from minutes to days. In this sense, the objective of this research is to analyze the performance of XGBoost model on solar irradiance forecasting and compare with the Tree Boosting model applied on the same forecasting situations. The XGBoost and Tree Boosting models were analyzed using different time scales (2 min, 10 min, 30 min, 1h, 1 day) from the same dataset that was partly used to train the models. This performance comparison was done using error metrics like RMSE, nRMSE, MAE, nMAE and others. The results were obtained using 50% of the original dataset, above of this percentage the values of errors do not have a significantly variation and allows to decrease the time of computing. Boosting model presents better performance than XGBoost in this kind of work, but with a slight similarity; XGBoost was faster to compute than the Boosting model. The performance of the models was affected by the ONI predictors in a positive way, where La Niña presents better FS values.*

Keywords: *solar irradiance forecasting, solar energy, machine learning, extreme gradient boosting, renewable energies.*

1. INTRODUCTION

Solar energy is one of the more extensive renewable resource in the world and is known as the source of the others renewable energies because it has important influence over them. Efficient uses of this source has been the biggest challenge for develop technologies to exploit it.

Intermittence is an aleatory characteristic of solar energy that depends on weather conditions, these conditions are variables that can't be controlled. Accurate prediction of solar irradiance is very important for successful integration with the power grid control system. (Kumari and Toshniwal, 2021). The complexity of modern power systems and the growth of on-grid integration with renewable energy demands higher management functionality for smart grids operation. It is difficult to develop forecasting tools with high accuracy using only one predicted value (Zhang *et al.*, 2018). There is no specific number of predictors to use in forecasting models and depends on which kind of model it is developed, it can be use endogenous or exogenous data.

Furthermore, the performance of solar forecast is affected by some factors as listed by Tuohy *et al.* (2015): look-ahead time as known as time horizon or timescale, variability of solar production, specific plant attributes, spatial scale and other weather phenomena. There are two well-known solar forecasting methodologies: cloud imagery compounded with physical models, and machine learning models; The first is used commonly for short-term forecasts and the latter is used for both short-term and long-term forecasts, in terms of timescale the accuracy is not the same (Voyant *et al.,* 2017).

Sharma *et al.* (2011) performed a comparison between machine learning and forecast-based models, results showed that Support Vector Machine – Radial Basis Function (SVM-RBF) with four of seven dimensions is 27% more accurate than a simple cloudy model and 51% more accurate than the Past-Predicts-Future (PPF) model. Even these forecast models included weather measurements but are unable to predict changes of weather patterns.

In order to investigate the impact of El Niño Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) it was developed an analysis by Davy and Troccoli (2012) using a bootstrap technique, the study indicated that ENSO phenomenon influences solar energy changes in more than 10% in some locations on a seasonal basis, inferring that forecasting models may use this variable as a predictor.

Gala *et al.* (2016) developed a comparison of Support Vector Machine (SVM), Gradient Boosted Regression (GBR), Random Forest Regression (RFR) and a hybrid method against a 3-h accumulated radiation forecast provided by Numerical Weather Prediction (NWP) systems for seven locations in Spain. They showed that the hybrid artificial intelligence systems have relevant better results for solar radiation forecast.

2. SCOPE OF THE PRESENT WORK

The present work aims to evaluate the performance of Boosting and XGBoost machine learning models as forecast tools to predict solar radiation, in addition to analyze the impact of La Niña and El Niño on the global solar radiation at the northeast city of Fortaleza using the Oceanic Niño Index (ONI).

3. METHODOLOGY

It is not previously determined a specific quantity of parameters to perform a forecasting model (Rocha *et al., 2019*). In order to develop this work, it was necessary to be used enough meteorological variables as well as the geographic location. Sections 3 and 4 show which predictors, details about the experimental dataset and machine learning models that were used to perform a regression model for solar radiation forecast. “Figure 1” shows the phases of the process developed in this work.

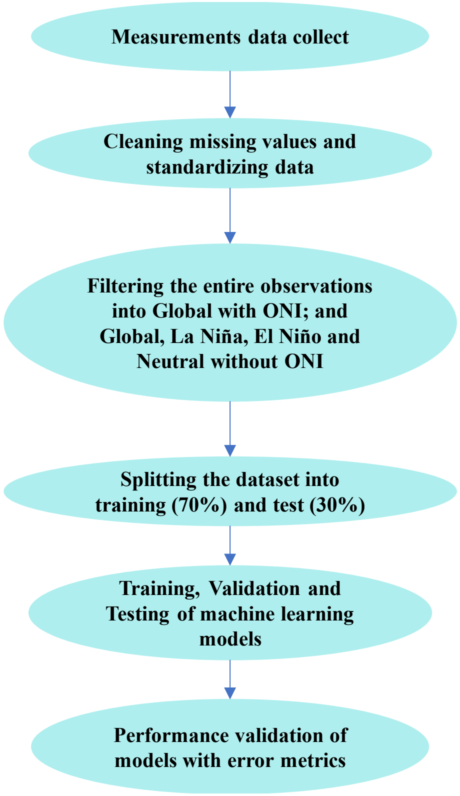
Measurements data collect
Cleaning missing values and standardizing data
Filtering the entire observations into Global with ONI; and Global, La Niña, El Niño and Neutral without ONI
Splitting the dataset into training (70%) and test (30%)
Training, Validation and Testing of machine learning models
Performance validation of models with error metrics

Figure 1. Flow diagram of the phases taken to build and test the models.

3.1 STUDY SITE AND EXPERIMENTAL DATA PRE-PROCESSING

The geographic location of the study site is in the Brazilian northeast city of Fortaleza, Ceará, 3° 43’ 2’’S, 38° 32’ 35’’O. Solar irradiance data was collected at Solar Energy and Natural Gas Laboratory (LESGN) in the Pici Campus of Federal University of Ceará and the meteorological measurements of temperature, wind velocity and direction, relative humidity and precipitation were obtained from Meteorology and Water Resources Foundation of Ceará (Funceme).

In addition to the predictors mentioned, the others used were year, day, time of measurements and irradiance of previous instants (2 min, 4 min, 6min, 8min,10 min, ..., 30min). The Oceanic Niño Index (ONI) is one of measurements of El Niño Southern Oscillation (ENSO) that was used too, it evaluates the intensity of the phenomenon with values from - 4 (Strong La Niña) to 4 (Strong El Niño) (CPC, 2005).

The dataset used involves the years from January 2007 to December 2013 (without 2009 and 2011), the variable day was standardized and the clearness index (Kt) was calculated as studied by Duffie and Beckman (2013).

4. MACHINE LEARNING MODELS AND ERROR METRICS

4.1 Persistence model

This is a trivial model that is normally used as the most common reference model in solar and wind short-term forecasting. Generally, it helps to benchmark other methods because the forecast accuracy decreases along the forecast duration for more than 1 hour ahead forecasting (Diagne *et al., 2013*). Equation (1) describes the persistence model, where the clearness index $\boldsymbol{K_t}$ at time $\boldsymbol{t+1}$ is predicted by its value at time $\boldsymbol{t}$,

$$K_{(t+1)} = K_t \tag{1}$$

4.2 Gradient Boosting model

This method is a meta-algorithm that combines an ensemble of weak decision tree learners and depth parameters which controls the number of binary evaluations allowed in each individual tree (Nagy *et al.*, 2016). The other parameter that works in the model is shrinkage, which operates controlling the weight of each decision tree (Landry *et al.*, 2016).

4.3 Extreme Gradient Boosting model

Among of the gradient boosting machine models the extreme gradient boosting model is the most popular of the tree algorithms. It works adding and training new trees to fit residual errors of the last iteration, and its advantages are minimal requirements for attributes normalization, processing missing values intelligently and greater performance for overfitting problems (Dong *et al.*, 2020). The XGBoost model is recognized by the most important characteristic that is scalability in all scenarios since it can run more than ten times faster than existing popular models (Chen and Guestrin, 2016). In the investigation of Fan et al. (2018), they show the general function (Eq. (2)) for the prediction at step *t*:

$$\boldsymbol{f}_i^{(t)} = \sum_{k=1}^{t} \boldsymbol{f}_k(\boldsymbol{x}_i) = \boldsymbol{f}_i^{(t-1)} + \boldsymbol{f}_t(\boldsymbol{x}_i) \tag{2}$$

Where the learner at step $\boldsymbol{t}$ is $\boldsymbol{f_t(x_i)}$; $\boldsymbol{f}_i^{(t)}$ and $\boldsymbol{f}_i^{(t-1)}$ are the predictions at steps $\boldsymbol{t}$ and $\boldsymbol{t-1}$, and $\boldsymbol{x_i}$ is the vector of input variables.

4.4 Error metrics for performance evaluation of the models

The performance of the models is evaluated using the following statistical indices: root mean squared error (RMSE) in Eq (3), normalized root mean squared error (nRMSE) in Eq (4), mean absolute error (MAE) in Eq (5), normalized mean absolute error (nMAE) in Eq (6). Also, to compare the performance of the models against a benchmark model, it is used forecast skill (FS) (Brasil *et al.,* 2020) given by the Eq (7).

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{\boldsymbol{N}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\left(\widehat{\boldsymbol{K}}_t - \boldsymbol{K}_t\right)^2} \tag{3}$$

$$\boldsymbol{nRMSE} = \frac{\boldsymbol{RMSE}}{\overline{\boldsymbol{K}}_t} \tag{4}$$

$$\boldsymbol{MAE} = \frac{1}{\boldsymbol{N}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\left|\widehat{\boldsymbol{K}}_t - \boldsymbol{K}_t\right| \tag{5}$$

$$\boldsymbol{nMAE} = \frac{\boldsymbol{MAE}}{\overline{\boldsymbol{K}}_t} \tag{6}$$

where $\widehat{\boldsymbol{K}}_t$ is the value forecasted by the model, $\boldsymbol{K}_t$ is the real observed value, $\boldsymbol{N}$ is the number of observations of the dataset and $\overline{\boldsymbol{K}}_t$ is the mean of $\boldsymbol{K}_t$ values.

$$\boldsymbol{FS} = 1 - \frac{\boldsymbol{RMSE}_{model}}{\boldsymbol{RMSE}_{persistence}} \tag{7}$$

The performance of the model is better as closer to 1 the $\boldsymbol{FS}$ value is.

5. RESULTS AND DISCUSSION

The analysis of the results comprehends three forecasting models working with five different sets; these are Global with ONI as predictor, that is the entire set of observations, as well as and the same Global dataset without ONI; La Niña years, El Niño years and Neutral years. These are composed by the years where happened each meteorological phenomenon, without the ONI predictor.

Table 1 features the results of Global data considering ONI as predictor used in the train process. It can noticed that the smallest $\boldsymbol{nRMSE}$ values are obtained from XGBoost model (11.81% - 2 min) and Boosting model (26.70% - 10 min; 30.03% - 30 min; 32.01% - 1 h; 41.73% -1 day). Likewise, Boosting model presents better $\boldsymbol{FS}$ values.

Table 1. Results of Global dataset with ONI.

Model	**Error metrics**	**Timescale**				
		2min	**10min**	**30min**	**1 h**	**1 d**
Persistence	RMSE	134.2216	185.9741	211.0821	224.4482	353.5532
	n.RMSE	23.38%	32.32%	36.62%	39.72%	44.29%
	MAE	68.7020	106.8340	136.0774	163.2498	273.1879
	n.MAE	11.97%	18.57%	23.61%	28.89%	34.22%
Boosting	RMSE	119.6337	153.6224	173.0984	180.8886	333.1274
	n.RMSE	20.84%	26.70%	30.03%	32.01%	41.73%
	MAE	67.7408	96.1347	115.8064	125.3577	279.5749
	n.MAE	11.80%	16.71%	20.09%	22.18%	35.02%
	FS	10.87%	17.40%	17.99%	19.41%	5.78%
XGBoost	RMSE	119.6225	154.4736	174.0291	184.7183	341.2422
	n.RMSE	11.81%	26.85%	30.19%	32.69%	42.75%
	MAE	67.7756	96.3710	115.4370	127.5722	277.0900
	n.MAE	11.81%	16.75%	20.03%	22.58%	34.71%
	FS	10.88%	16.94%	17.55%	17.70%	3.48%

Table 2 features results of Global dataset not considering ONI as predictor. In general, the values of $\boldsymbol{nRMSE}$ are slightly higher than the values of Table 1, what means that the inclusion of ONI as part of the predictors impacts the performance of the models.

One can observe that values of $\boldsymbol{RMSE}$ increase along the timescale. The better values of $\boldsymbol{MAE}$ in general from Table 1 and Table 2 were obtained by XGBoost (67.77 – 2 min) and Boosting (67.74 – 2 min) considering ONI. As can be noted, the values are comparable, remembering that XGBoost model works faster than the other model.

Table 2. Results of Global dataset without ONI.

Model	**Error metrics**	**Timescale**				
		2min	**10min**	**30min**	**1 h**	**1 d**
Persistence	RMSE	133.9211	187.1050	200.8503	220.4350	386.3919
	n.RMSE	23.30%	32.67%	35.18%	39.30%	48.21%
	MAE	69.0152	106.6954	128.5657	161.6070	303.7206
	n.MAE	12.01%	18.63%	22.52%	28.81%	37.90%
Boosting	RMSE	120.0591	154.4526	166.4855	177.9720	331.2364
	n.RMSE	20.89%	26.97%	29.16%	31.73%	41.33%
	MAE	68.0428	95.9546	108.1667	127.1401	280.2253
	n.MAE	11.84%	16.75%	18.95%	22.66%	34.96%
	FS	10.35%	17.45%	17.11%	19.26%	14.27%
XGBoost	RMSE	120.0907	154.4739	168.6203	180.7548	338.6287
	n.RMSE	20.89%	26.97%	29.54%	32.22%	42.25%
	MAE	68.1032	96.1367	109.8204	127.3951	282.8027
	n.MAE	11.85%	16.78%	19.24%	22.71%	35.29%
	FS	10.33%	17.44%	16.05%	18.00%	12.36%

As can be seen in the Figure 2 and Figure 3, the ***FS*** have a better performance in the analysis that uses ONI, increasing along the timescale up to 60 min and remarkably decreasing after that.

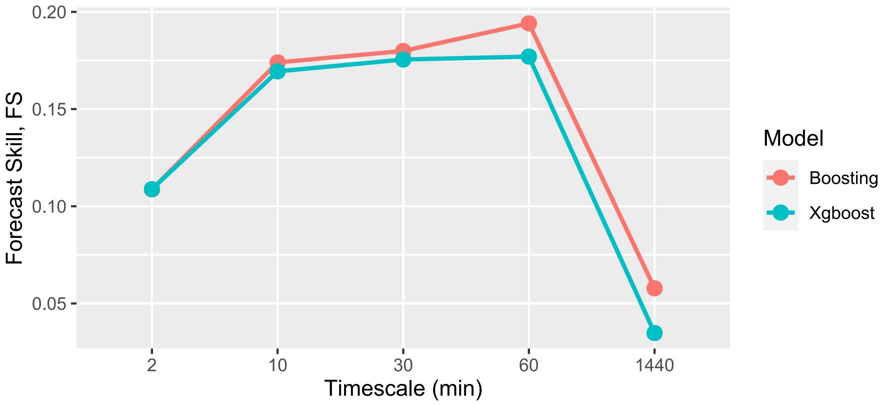

Figure 2. Forecast skill of Global dataset with ONI.

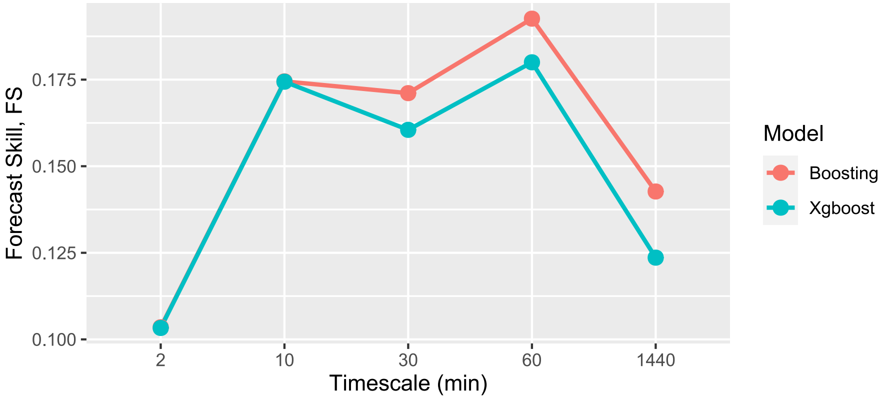

Figure 3. Forecast skill of Global dataset without ONI.

Figure 4 features the values of ***FS*** obtained with the set La Niña, which do not consider ONI. It can be noticed the improved values of ***FS*** comparing with Figure 5 and Figure 6. The five studied datasets show similar trends, decreasing ***FS*** from 1 h to 1 d timescale, but only the XGBoost model shows a negative ***FS*** (-6.55%) for the Neutral years dataset in a 1 d timescale.

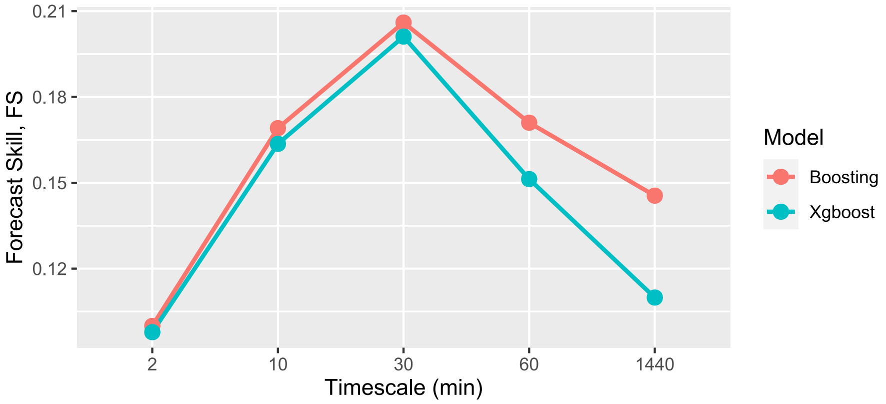

Figure 4. Forecast skill of La Niña years dataset without ONI.

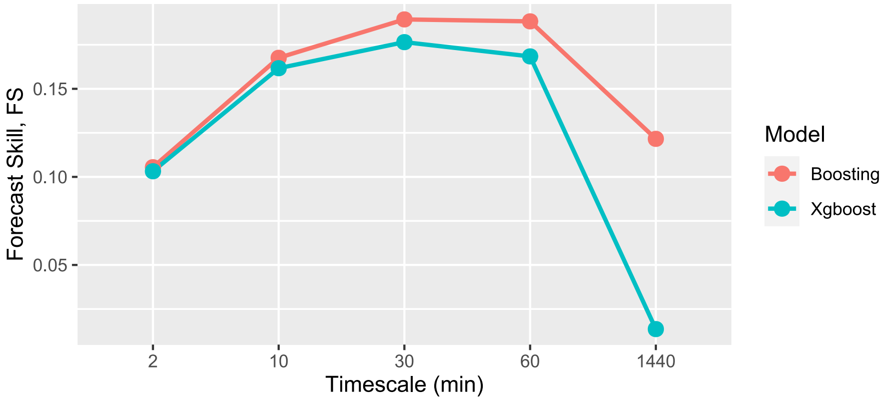

Figure 5. Forecast skill of El Niño years dataset without ONI.

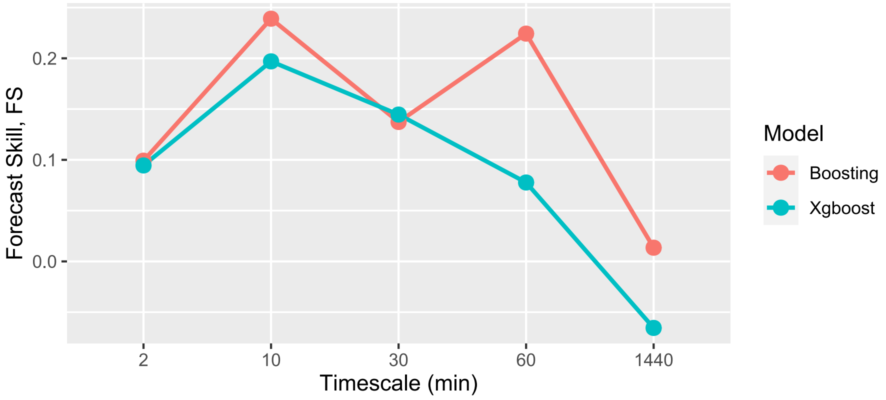

Figure 6. Forecast skill of Neutral years dataset without ONI.

Comparing the dataset Global with ONI against the separated datasets La Niña, El Niño and Neutral years which do not use ONI as predictor, it can be noticed that the datasets which were divided in years when occurred or not the ENSO got higher values of ***FS***. This approach can then be indicated to improve the performance of solar forecasting using historical data of ENSO years.

6. CONCLUSION

This work was developed with meteorological data that involves years from January 2007 to December 2013 (without 2009 and 2011). This data was pre-processed and standardized avoiding missing values and setting properly the types of variables according to the input parameters of each model.

It was accomplished a pre-training/testing of the models using different percentages of dataset (10%, 20%, 30%, 40%, 50%), because the total observations are 512267, and it is quite large to handled. With this analysis it was possible to know that results do not have a significant variation in general, then is not necessary work with the entire dataset. This analysis was made apart of and before the filtering and splitting sets.

The performance of both models was slightly similar, and the Boosting model presents better values, but the XGBoost model performs not too far. In this moment it is important to highlight the computing speed of XGBoost model, which was notably faster than Boosting in the process of Global sets which have more observations than the others ones.

In cases where it was used the ONI predictor, it can be noticed that exists a positive influence over the performance of models, as well as with the separations of the sets of years where take place the meteorological events La Niña and El Niño.

7. ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)— Finance Code 001, and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazilian governmental agencies. The support received is gratefully acknowledged.

8. REFERENCES

Brasil, J. S., Marinho, F.P., Costa Rocha, P.A., 2020. “Influence of el niño and la niña on the intrahorary forecast of horizontal global radiation”. *Brazilian Journal of Development,* Vol. 6, No. 1, pp. 2321-2329.

Chen, T. and Guestrin, C, 2016. “XGBoost: A Scalable Tree Boosting System”. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)”. *Association for Computing Machinery*, New York, NY, USA, pp. 785–794.

Costa Rocha, P.A., Fernandes, J.L., Modolo, A.B., Pontes Lima, R.J., Vieira da Silva, M.E., Bezerra, C.A., 2019. “Estimation of daily and monthly global solar radiation using ANNs and a long data set: a case of study of Fortaleza, in Brazilian Northeast region”. *International Journal of Energy and Environmental Engineering,* Vol. 10, No. 3, pp. 319-334.

CPC, 2005. “Historical El Nino / La Nina episodes (1950-present)”. *Climate Prediction Center, National Oceanic Atmospheric Administration*. https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php. Acessed em: 17 March 2021.

Davy, R.J., Troccoli, A., 2012. “Interannual variability of solar energy Generation on Australia”. *Solar Energy,* Vol. 86, No. 12, pp. 3554-3560.

Diagne, M., David, M., Lauret, P., Boland, J., & Schmutz, N. 2013. “Review of solar irradiance forecasting methods and a proposition for small-scale insular grids”. *Renewable and Sustainable Energy Reviews*, Vol. *27*, pp. 65–76.

Dong, W., Huang, Y., Lehane, B., Ma, G., 2020. “XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring”. *Automation in Construction,* Vol. 114, Art. No. 103155.

Duffie, J.A., Beckman, W.A., 2013. “Solar Engineering of Thermal Processes (in english)”. John Wiley & Sons, Hoboken, New Jersey.

Fan, J., Wang, X., Wu, L., Hanmi, Z., Zhang, F., Yu, X., Lu, X., Xiang, Y., 2018. “Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China”. *Energy Conversion and Management,* Vol. 164, pp. 102-111.

Gala, Y., Fernández, Á., Díaz, J., & Dorronsoro, J. R. 2016. “Hybrid machine learning forecasting of solar radiation values”. *Neurocomputing*, Vol. *176*, pp. 48–59.

Kumari, P., Toshniwal, D., 2021. “Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance”. *Journal of Cleaner Production,* Vol. 279, Art. No. 123285.

Landry, M., Erlinger, T.P., Patschke, D., Varrichio, C., 2016. “Probabilistic gradient boosting machines for GEFCom2014 wind forecasting”. *International Journal of Forecasting,* Vol. 32, No. 3, pp. 1061-1066.

Nagy, G.I., Barta, G., Kazi, S., Borbély, G., 2016. “GEFCom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach”. *International Journal of Forecasting,* Vol. 32, No. 3, pp. 1087-1093.

Sharma, N., Sharma, P., Irwin, D., & Shenoy, P. 2011. "Predicting solar generation from weather forecasts using machine learning". *2011 IEEE International Conference on Smart Grid Communications, SmartGridComm 2011*, pp. 528–533.

Tuohy, A., Zack, J., Haupt, S. E., Sharp, J., Ahlstrom, M., Dise, S., Grimit, E., Mohrlen, C., Lange, M., Casado, M. G., Black, J., Marquis, M., & Collier, C. 2015. "Solar Forecasting: Methods, Challenges, and Performance". *IEEE Power and Energy Magazine*, Vol. 13, No. 6, pp. 50–59.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A., 2017. "Machine learning methods for solar radiation forecasting: A review". *Renewable Energy*, Vol. 105, pp. 569–582.

Zhang, W., Quan, H., Srinivasan, D., 2018. "Parallel and reliable probabilistic load forecasting via quantile regression forest and wuantile determination". *Energy,* Vol. 160, pp. 810-819.

9. RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.