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

COB-2021-0247
NOWCASTING AND SHORT TERM GHI FORECASTING USING GOES-16 SHORTWAVE RADIANCE DATA: A MACHINE LEARNING CASE STUDY OF PETROLINA-PE, BRAZIL

Paulo Alexandre Costa Rocha
Solar Energy and Natural Gas Laboratory, Mechanical Engineering Department, Technology Center, Federal University of Ceará, Fortaleza, CE 60020-181, Brazil
paulo.rocha@ufc.br

Hugo T. C. Pedro
hpedro@eng.ucsd.edu
Carlos F. M. Coimbra
Department of Mechanical and Aerospace Engineering and Center for Energy Research, University of California San Diego, La Jolla, California 92093, USA
ccoimbra@ucsd.edu

Abstract. *In this work it was used the visible and near-IR channels of the GOES-16 satellite to model (nowcasting) and forecast GHI intra-hourly. The signals were used as predictors, both in their raw and normalized forms. Five different machine learning models were tested, namely Artificial Neural Networks, k-Nearest Neighbors, LASSO, Support Vector Machines and XGBoost. Their performance was compared by the RMSE and Forecast Skill relative to the persistence model. The first four methods presented similar behavior, with the RMSE values ranging from 180.42 to 202.05 W.m^{-2}. The XGBoost outperformed all the other methods, obtaining RMSE values between 144.37 and 160.34 W.m^{-2}. The FS improved over the time horizon for all methods, being negative for nowcasting and 15-min forecasting, with the exception of XGBoost, which performed positively for all forecasting time horizons, reaching 40.78 % for 60-min horizons.*

Keywords: *renewable energy, solar irradiance forecasting, full disk GOES-16 satellite, machine learning, caret R package*

1. INTRODUCTION

Solar energy harvesting technologies have been developing consistently in a world that is increasingly looking for options for the sustainability of its activities. There is already clear evidence of the link between the use of renewable technologies and economic growth (Saidi and Omri, 2020), in addition to a regular growth in installed capacity, which in 2018 reached 2,356.35 GW (IRE, 2020). In this context, solar energy has taken a prominent role as a clean energy source.

Brazil has been steadily increasing the use of solar energy, which is still small, but should follow the world trend in the near future. Specifically for photovoltaic (PV) technology, between January 2018 and January 2019 there was an increase of 115.2 % in the Brazilian energy matrix (de Minas e Energia, 2020).

Thus, knowledge of the characteristics of this source has been shown to be important, especially its intermittent features, which have been the subject of studies for several decades (Chow *et al.*, 2011). These studies highlight the relevance of short-term forecast, both operationally and economically (Pedro and Coimbra, 2012). Still, given the relevance and complexity of the topic, the research has accelerated in recent years focusing on aspects such as new forecasting algorithms (Caldas and Alonso-Suárez, 2019) or the choice of the clear-sky model normalization (Yang, 2020).

As in any forecasting problems the input variables have significant importance in the models' accuracy. In this sense, there are basically three common approaches when doing solar nowcasting and forecasting: endogenous, where models use previous measurements of the target variables and/or values from nearby sensors; exogenous, where the model uses explanatory variables other than the target; and hybrid, which is the usual approach, that uses endogenous and exogenous data such as meteorological variables (Rocha *et al.*, 2019), sky images (Feng and Zhang, 2020) or satellite data (Zagouras *et al.*, 2015).

This work uses pure exogenous variables, consisting in data of the six shortwave channels from the GOES-16 to nowcast and forecast GHI. The *Geostationary Operational Environmental Satellite* (GOES) is a set of meteorological satellites operated by the *National Oceanic and Atmospheric Administration* (NOAA), where the GOES-16, whose images were used in this research, is the first of their new generation. This allows to assess the solar radiance at any point covered by the satellite, without the necessity of a ground instrumentation.

Satellite data have been the object of study in the field of solar energy (Yang and Bright, 2020). They carry information for the determination and/or prediction of irradiation, both using physical based (Larson *et al.*, 2020; Kallio-Myers *et al.*, 2020) and statistical based models (Cornejo-Bueno *et al.*, 2019). In these works RMSE values can be found in the range of 107.19-121.63 W.m^{-2} for hourly-average GHI nowcasting, when only satellite data was used as the models inputs.

In our case, a narrower average of 15 minutes was used, which leads to larger RMSE values due to the intrinsic larger GHI variability.

Machine learning (ML) models have been applied frequently in studies regarding renewable energy sources and power generation. Some examples in the literature include research related to biomass energy production (Elmaz *et al.*, 2020), wind energy, both for forecasting (Demolli *et al.*, 2019) as well as for characteristics of turbine operation (de Abreu [Melo Junior] *et al.*, 2019). Regarding methods such as *k*-Nearest Neighbors (*k*-NN) (Inman *et al.*, 2013), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN) (Zhao *et al.*, 2019), Support Vector Machines (SVM) (Lima *et al.*, 2020) and Tree-Based methods (Fan *et al.*, 2020), among several others have been applied. Specifically, this study assesses the performance of five ML methods, namely: Least Absolute Shrinkage and Selection Operator (LASSO), *k*-NN, ANN, SVM and eXtreme Gradient Boosting (XGBoost). They are used for both nowcasting and forecasting for time horizons of 15, 30, 45 and 60 minutes.

The main contributions of this research effort are: the use of the new GOES-16 satellite data for the GHI evaluation using ML; the performance comparison of several ML methods for that task; and the application, tuning and evaluation of the XGBoost with satellite data in solar irradiation nowcasting and forecasting. Furthermore, we consider noteworthy that the studied spot, Petrolina city in Brazil, needed the *Full Disk* GOES-16 data. Its position lies quite under the satellite nadir (near the Equator), and we could not find research under these premises.

After this introduction, the work is structured in the following manner: The data used in this research are described in Section 2, both the ground based data (Sec. 2.1) and the satellite data (Sec. 2.2); the data treatment is described in Section 2.3, including data normalization, while preprocessing algorithms are explained in Section 2.4; the ML methods are presented (Sec. 2.5), and their results are shown and discussed (Sec. 3), and their performance compared in Section 3.6. Finally, the conclusion is presented (Sec. 4).

2. DESCRIPTION OF THE DATA

The data used in this research came from two sources. Ground measurements of GHI were obtained from the Brazilian project *Sistema de Organização Nacional de Dados Ambientais* - Environmental Data National Organization System (SONDA) (SON, 2020). GOES-16 remote sensing data were obtained from the *Google Cloud Platform* (Platform, 2020), via the *Google Software Development Kit* (SDK) (Google, 2020).

2.1 Petrolina GHI Data

The SONDA network was created to implement a physical and human resources infrastructure in order to raise and improve the database of solar and wind energy resources in Brazil. It has 20 locations spread across the country, including radiometric, anemometric and hybrid stations. One of the advantages for selecting these data is the good control in the measurements validation. The city of Petrolina, object of this study, was selected because of its characteristics, being one of the places with the highest values of GHI in Brazil, and also because it is close to the Equator, thus close to the center of the image generated by GOES-16. As already mentioned, the full disk image had to be used, what is not found in the literature, where the continental USA (CONUS) is the common approach.

The data used are from the year 2018. They are originally acquired in 1-minute averages, and validated based on the data quality control strategy adopted by the *Baseline Surface Radiation Network* (BSRN). Because of the satellite temporal resolution, 15 minutes averages were taken and used in the modeling. Despite the lack of data in some periods, notably for the months of August to November, the fact that they are consistently validated counts positively for confidence in the results presented.

2.2 GOES-16 Data

GOES-16 is a satellite that has advanced Earth's imagery and atmospheric measurements of Earth's weather, practically in real-time, among other relevant features (NASA, 2020). It was launched in 19/Nov/2016, officially declared GOES-East in 18/Dec/2017, and its first image was acquired and processed in 17/Jan/2017.

One of GOES-16 most important resources for weather monitoring and forecasting is the Advanced Baseline Imager (ABI), which is capable of viewing the Earth in 16 different spectral bands, far more than the 5 available in the previous GOES-East. ABI captures images at three different scales: mesoscale, CONUS and full disk. The first tracks small areas of interest every minute (e.g. storms), the second captures continental USA (CONUS) every 5 minutes and the third captures the full disk. In this case, Petrolina is only visible in the full disk image which is available every 15 minutes (in April 2019 ABI scan mode was changed to enable a full disk 10-minute scan).

In this work, the first six bands, which account for the shortwave spectrum, were used for the now- and forecasting of GHI. The radiance values of the channels for the pixel in which the Petrolina station is located were taken as inputs, together with the other predictors presented in Section 2.3 Furthermore, the ABI has in orbit calibration capacity, which is detailed in Kalluri *et al.* (2018). The article also describes image scanning, the internal calibration procedure, and several interesting characteristics regarding the time and spatial resolutions.

2.3 Data Treatment

As previously mentioned (Sec. 2.2), this work uses pixel-wise data from channels 1 to 6, which represent the measured shortwave radiance in W.m^{-2}.sr.cm^{-1}. It should be noted that, for a, e.g. 2 km^2 pixel, the reported radiance is not the average of the 2 km $\times$ 2 km area, but the best estimate of the radiance in its center (Kalluri *et al.*, 2018).

The predictors used in the models include the current pixel-wise value for each channel and their corresponding previous 15-, 30-, 45- and 60-minute values, comprising a total of 30 radiance values. In addition, the day of the year, the minute of the day, as well as the solar zenith angle were used, totaling 33 predictors for each simulation.

Two types of models are generated in this work depending on data treatment. The first uses the raw data without any normalization or detrending. The second uses normalized radiance values following Eq. 1.

$$x_{norm} = \frac{x_{measured}}{x_{ref} \cos(\theta_z)} \tag{1}$$

where x_{norm} is the normalized variable, which resulted from the transformation of $x_{measured}$ over x_{ref} and the detrending term $\cos(\theta_z)$. For the GOES-16 data, the x_{ref} is the maximum value of each band.

It is a common practice in ML to normalize variables, since some algorithms tend to work better with variables on the same scale of magnitude and centralized, approaching a normal distribution as much as possible.

In our normalization approach, the difference between scales is narrowed, but the distribution of values is not centralized in relation to the origin. Scaling and centering (implemented with the *caret* package) were also tested, and kept whenever improving the results. Details can be found in Section 2.4 Additionally, it is a common practice in autocorrelated time series, such as solar irradiance, to apply a detrending transformation using the diurnal and seasonal cycle, usually via the solar zenith angle.

For the case of the target variables, these consisted of the GHI values, at time t in the case of nowcasting, and at times $t + \Delta t$ ($\Delta t = \{15, 30, 45, 60\}$ minutes) for intra-hour forecasts. For simulations with normalized values, the target variables also went through this process with $x_{ref} = 400$ Wm^{-2}, which is considered a reasonable approximation of the average GHI at Petrolina.

For all the applied models in this study, a unique separation of the training set occurred at a random 70% - 30% of the original data base. Although common in the literature, chronologically sequential 70% - 30% separation was not used in order to include data from all seasons in the training and testing sets.

2.4 Predictors Preprocessing

Several preprocessing techniques were used on the predictors. For specific models, they can improve their performance. The *caret* package comes with a set of preprocessing tools available for automatic use. Table 1 presents the predictors preprocessing techniques that were tested for all methods. For each method, the approaches that improved the results are marked. Here it is worth to mention that, even though the table is organized by method, for every time horizon the preprocessing tools were tested separately. It was found that the best configurations for a given ML model did not change with the forecasting horizon. For the LASSO method, no predictor preprocessing was available in the library.

In the ANN case, *center* and *scale* are common approaches, which set the predictors to values between -1 and 1. They transpose the predictor values into the data range that the activation functions usually lie.

Further preprocessing techniques (for both raw and normalized variables) were also employed. These include Principal Components Analysis (PCA) (Pearson, 1901; Hotelling, 1936) and *Spatial Sign* (Serneels *et al.*, 2006). *Spatial Sign* consists in projecting each predictor vector to the unit circle in *p* dimensions, where *p* is the number of predictors. In a direct way, each vector of data related to each predictor is divided by its norm. In the case of the raw variables, PCA presented the best results, and for the normalized variables *Spatial Sign* performed better. This may be due to the characteristics of the raw predictors. They show a cyclic behavior which increases the correlation between the satellite signals, and consequently the model variance. PCA tends to reduce the predictors correlation, imposing a new and smaller set of low correlated variables. Besides improving the results, in *caret* implementation of ANN, PCA has, as a drawback, the loss of physical insight of the predictors, since as part of the process, the principal components are created to substitute the original variables. Nevertheless, one advantage is the dimensional reduction, which reduces the computational time considerably.

In the *k*-NN implementation, no predictor algorithm preprocessing was used, but *scale*, *center*, *spatial sign* and *pca* were tested, separately and in combinations. As already mentioned, the same approach was applied to all ML methods of

this work.

Regarding the SVM ML model, the predictors preprocessing was done with *center*, *scale* and *pca*, which was the configuration that presented the best results. The *spatial sign* preprocessing procedure was tested, with and without *pca* together, with no improvement in the results, for the case of raw variables. For the normalized ones, *spatial sign* improved the results of RMSE, which did not happen by merging *spatial sign* and *pca*. This was the same preprocessing configuration of ANN, and the reasons for this, shown in Section 3.1, may be the same as well.

For the XGBoost model, in the case of the raw variables approach, none of the tested preprocessing methods improved the results, while for the normalized variables, the *center* approach did.

Table 1. Variables preprocessing tested for all the studied models. The approaches that presented the best RMSE values are marked. For the LASSO method, no predictors preprocessing was applied.

	Raw Variables				Normalized Variables			
	Center	Scale	PCA	SpatialSign	Center	Scale	PCA	SpatialSign
ANN	X	X	X		X	X		X
k-NN								
SVM	X	X	X		X	X		X
LASSO	-	-	-	-	-	-	-	-
XGBoost					X			

2.5 ML Models

One of the objectives of this research is to evaluate the performance of several ML methods in the now- and forecasting of GHI values, using only satellite data and date and time features. Machine Learning is a pure statistical-computational approach, and in this sense no physical insight is passed to the models, which are trained to find the dependence between the input variables, as far as their intrinsic importance to the results. Five different ML methods were tested, trying to cover some desirable capabilities of now- and forecasting. These capabilities are robustness, non-linearity (*k*-NN, ANN, SVM, XGBoost) and variable selection (LASSO, XGBoost). A dull persistence model (Eq. 2) was also used as a comparison standard of the models performance.

$$\hat{I}\left(t+\Delta t\right)=I\left(t\right) \tag{2}$$

where $\hat{I}\left(t+\Delta t\right)$ is the GHI value predicted by the immediate predecessor GHI value under a Δt time horizon.

The *caret* package (Kuhn, 2008) in *R* language was used to implement all models but for LASSO. In this case, we used the *glmnet* package also in *R*.

To compare the models performance, the *Root Mean Squared Error* (RMSE), Eq. 3, and the *Forecast Skill* (FS) Eq. 4, were evaluated. The FS indicates how better or worse, in percentage terms, a model performs in comparison to the reference persistence model, being negative when worse, and positive when better. The FS has been recommended in the recent literature (Yang *et al.*, 2018) as a standard to be adopted in solar forecasting.

$$\text{RMSE}_{model}=\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\hat{I}(t)-I(t)\right)^{2}} \tag{3}$$

$$FS=1-\frac{\text{RMSE}_{model}}{\text{RMSE}_{persistence}} \tag{4}$$

3. RESULTS

This section presents the results obtained in the simulations, for all the time horizons studied. We begin by presenting the results of the each model separately; in the end we present a comparison between all models, with the evaluation of their performance after the tuning procedure. All presented results are for the hold-out testing set. As previously presented (Sec. 2.3), the training and testing sets are the same for all models, and they correspond to 70% and 30% of the original data set, respectively.

3.1 Artificial Neural Networks (ANN)

ANNs have been used for a long time in solar resource forecasting (Pazikadin *et al.*, 2020). In this work, the *nnet* method of the *caret* package was used to implement an ANN with a single hidden layer.

Regarding hyperparameter tuning we tested the weight decay and the number of neurons. This experiment yielded optimal values between 0.2-0.9 for the weight decay and 15-29 for the number of neurons. These hyperparameter intervals did not change for both approaches using raw and normalized variables.

Values of RMSE ranged from 182.87 W.m^{-2}, for the GHI nowcasting using normalized variables and 202.05 W.m^{-2} for the nowcasting using raw variables. To qualitatively illustrate the results obtained, Fig. 1 presents the dispersion between the measured values and those predicted by the model. Plot (a) (left) refers to the simulation using raw variables as predictors, while plot (b) (right) illustrates the results using normalized variables. It is noteworthy the superiority of the results with the normalized variables. They show smaller dispersion, and the absence of repetitive values of GHI as seen in plot (a) (evidenced by the subtle horizontal lines formed by the overlapping of repeated results).

Regarding the comparison between the measured and predicted GHI histograms, there is no significant difference between raw variables and the normalized ones. This confirms that the proposed transformation did not statistically distort the response variable. The predicted GHI histograms have a different shape, without a marked peak for low values, and with a peak for values between 500 and 700 W.m^{-2}. The scatter plot also indicates that the model tends to overestimate the predictions.

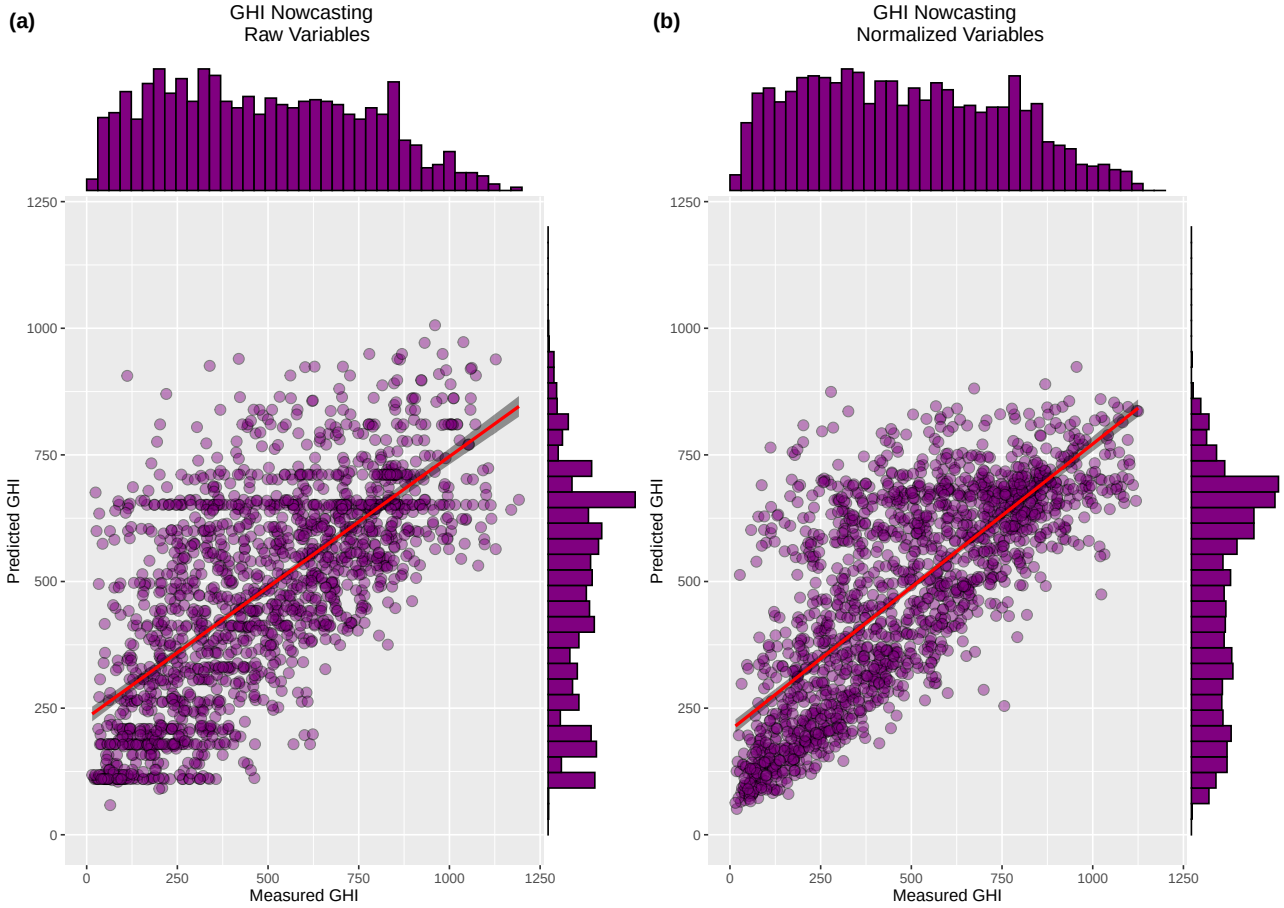

Figure 1. Comparison of the predicted results for the GHI nowcasting using ANN. Both graphs (a) and (b) present a regression line in red to improve the visualization of the predictions tendency. Graph (a), in the left side, presents the results for the raw variables model implementation. Graph (b) regards the same for the normalized variables. The histograms of the testing set are presented on the top, and the histograms of the predicted GHI set lies on the right of each graph.

Regarding the results of the model tuning, for both for the raw variables and the normalized ones, a grid of hyperparameter values is created and the model is trained using the 10-fold cross validation (James *et al.*, 2013). Several runs using different subgrids are required to find optimal values, by a trial and error procedure, always following the values of the parameters associated to the best results from the previous grid. This approach tends to be faster than creating a unique large grid, which may waste computational time on combinations of parameters that will not result in low RMSE values. For ANN, the RMSE variation is small, ranging from $\approx$3 % for the raw variables and 1 % for the normalized case, indicating that a fine tuning was reached.

An analysis of the variables importance was performed and the results are shown (Fig. 2), for the case of 30 min ahead forecasting. As already stated, there was a loss of information from the predictors for the raw variables (a), where PCA

was used to improve the model. Even so, one can observe the dimensional reduction, where out of 33 initial predictors, 10 principal components were generated. In the case of normalized variables (plot (b) in Fig. 2), the three related diurnal and seasonal cycles had a greater impact on the results, as expected. The information of the instantaneous reading of each channel proved to be important, as well as the lagged data of 15 and 30 minutes. Radiance values older than 45 and 60 minutes were practically irrelevant to the model.

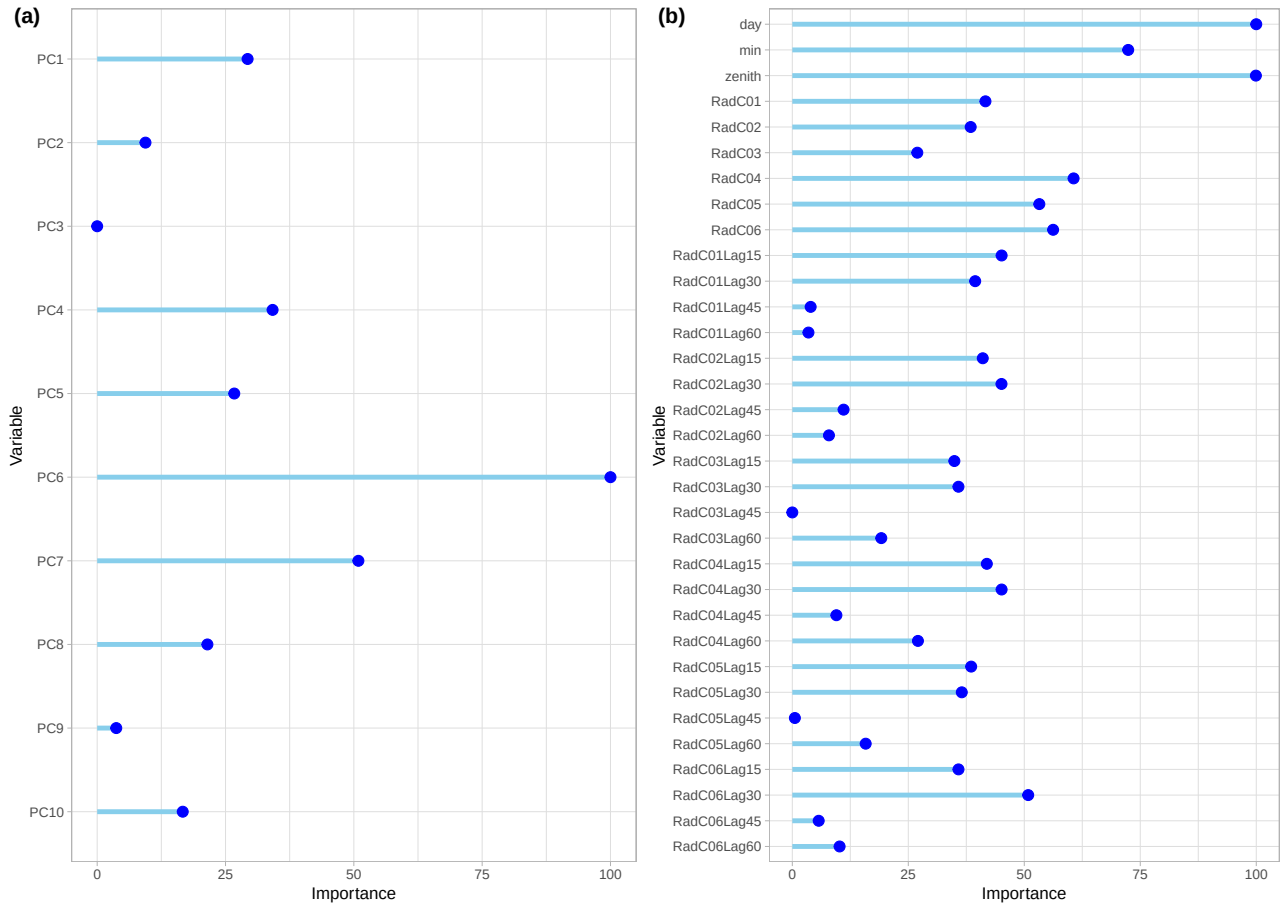

Figure 2. Plot of variable importance in the 30 minutes forecasting of the ANN ML method, for the raw variables case (a) and normalized variables case (b). In the left plot (a), the original variables were substituted by the principal components, what did not happen in the right plot (b), in which PCA was not employed.

3.2 *k*-Nearest Neighbors (*k*-NN)

The *k*-Nearest Neighbors algorithm (*k*-NN) is a classical ML method, which can be used as a performance reference for other methods. The *k*-NN is based on the estimation of the response function through a measure of proximity to the neighboring values.

In this work, the *knn* method of *caret* was used. A grid search for the number of neighbors *k* (Fig. 3) resulted in optimal values in the range of 15-25, with a majority of values less than or equal to 20. The most notorious case was for 45 minutes forecasting of normalized variables, where k = 28 and the convergence was oscillatory, so this case was chosen to be featured here. High *k* values did not impact performance, as *k*-NN has proven to be robust and lightweight for virtually any current computing platform

Regarding the analysis of the cross comparison between the measured values and those predicted by *k*-NN, once again there was a tendency towards overestimation, with a little more dispersion in the case of raw variables. Histograms also showed divergence, with a relatively small amount of low GHI values. For this case, 45 minutes ahead, the RMSE values were 187.74 W.m^{-2} and 182.79 W.m^{-2}.

As well as for the ANN, the evaluation of the variables importance was done, and it can be seen the effect that the variables normalization has on the variables measured by the satellite. The information provided gains relevance, in addition to making more evident which channels were important for the model. The current value of channel 4 has virtually no impact on the model, which may be because its information is already embedded in the values of the other channels. Furthermore, over the time horizon, the lags of 15 and 30 minutes were relevant for practically all channels,

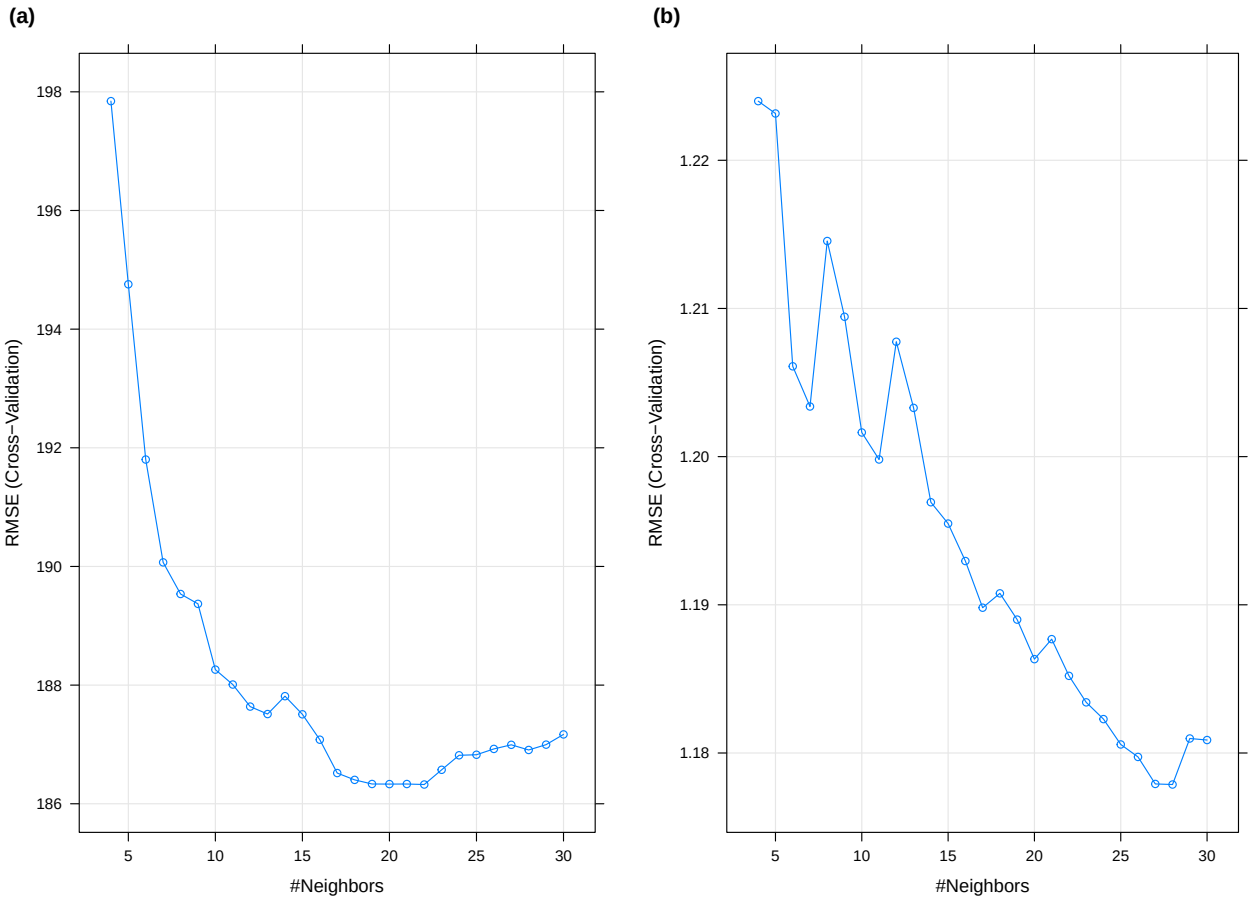

Figure 3. Example of *k* convergence plot for the GHI 45 minutes forecasting, both for raw variables (a) and normalized variables (b) where, in this specific case, the convergence occurred in an oscillatory but consistent way.

except again for channel 4.

3.3 Support Vector Regression (SVR)

Originally designed for binary classification, SVM can be adapted for classification of more than two classes, as well as for regression, and in this case it is also called Support Vector Regression (SVR) (Ayodele *et al.*, 2019).

SVR models of the *caret* package were tested with linear (*svmLinear*), polynomial (*svmPoly*), exponential (*svmExpoString*) and radial kernels with sigma adjustment (*svmRadialSigma*). Several values of the σ and C hyperparameters were tested, but it was found that the default values automatically generated by *caret* showed better performance. In this case σ is calculated by the function *sigest* from the *kernlab* package. One can also see an inversion in the overfitting trend in relation to raw and normalized variables. The growth of σ in the first case may imply a mismatch with the training data, the opposite occurring in the second. Since the 10-fold approach was used to optimize the parameters, it is expected that this trend will be replicated in the test set.

It is worth mentioning here that even using PCA, for SVM differently from ANN, *caret* keeps the name of the variables, not in main components, which facilitates the analysis of their importance. Thus, it can be seen in Fig. 4 that the normalization of the variables was once again able to increase the influence of the satellite data, even with the variable *zenith* still remaining the most important for the model. Regarding the time lagging, there is also a greater relative importance of the values for 15 and 30 minutes, practically for all channels.

In the simulation results using SVM, it still indicates the overestimation of GHI values, with a slightly greater dispersion below the regression line for the normalized variables (b). Histograms demonstrate this, with a more dispersed peak around the GHI value of 750 W.m^{-2}. This uniformity can also be noticed by the RMSE and FS values, which were slightly better in the normalized case (RMSE=189.05 W.m^{-2}, FS=25.29 %) in comparison to the raw variables case (RMSE=190.50 W.m^{-2}, FS=24.72 %).

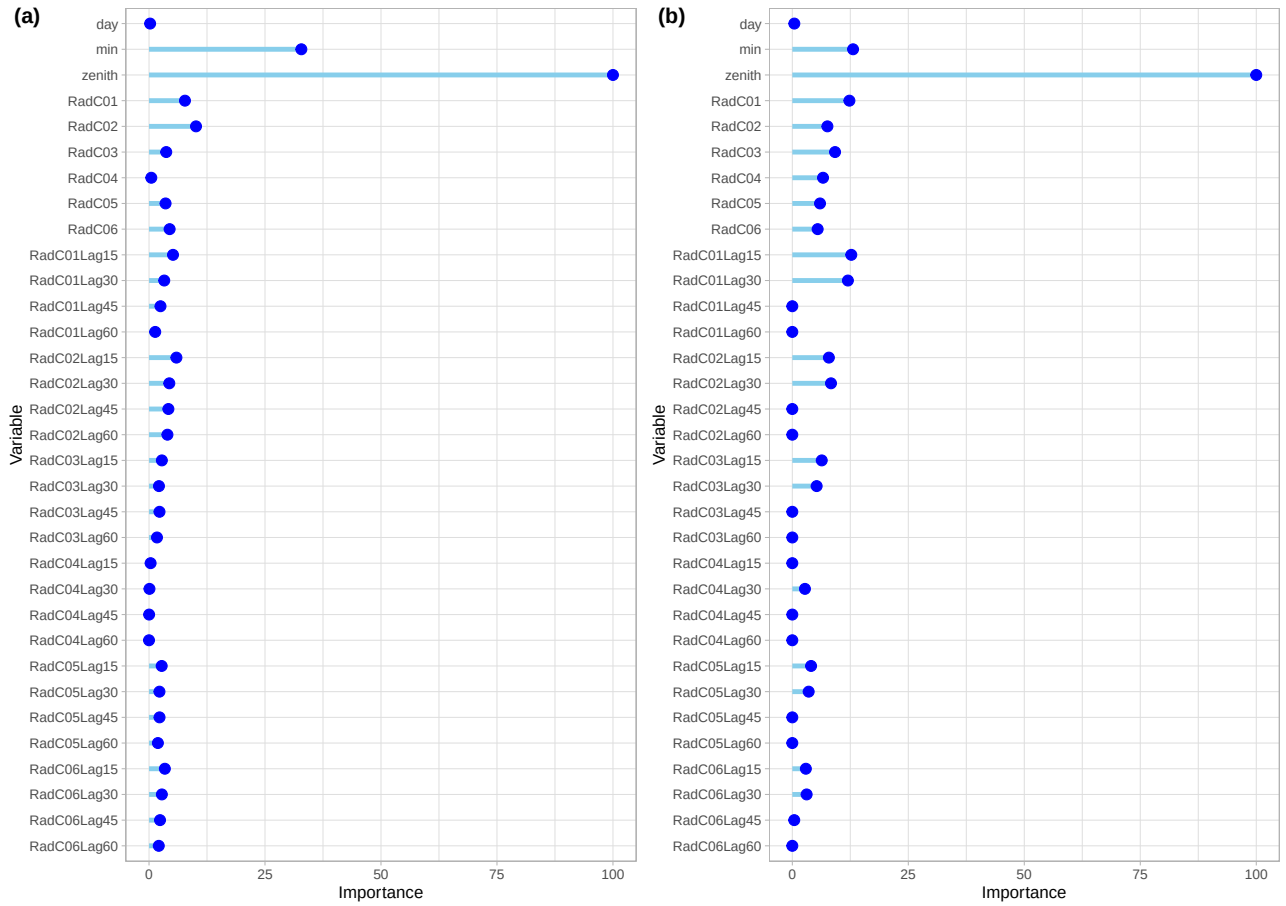

Figure 4. Same as Fig. 2, for the 60 minutes forecasting of the SVM ML method. In the normalized variables case (b), the influence of channel 4 could be captured by the model.

3.4 Least Absolute Shrinkage and Selection Operator (LASSO)

(LASSO) is a linear regression method, where a penalty term is applied to solve the optimization problem, reducing the value of the absolute sum of the equation coefficients, with an approach similar to *Ridge Regression*. This penalty firstly tends to reduce the coefficients of the variables that have the least impact on the optimal solution (Tibshirani, 1996).

Because of its efficiency and ease in the interpretation of results, the LASSO is applied in many practical problems, including solar energy (Yang *et al.*, 2015), where the spatial selection of the most important measurement points for GHI naturally happens with the application of the method. In ML there is no free lunch (Wolpert and Macready, 1997) though, so what is gained in interpretability is lost by the fact that LASSO is purely linear. Thus, to insert nonlinearity characteristics in the model, polynomials of degrees from 1 to 7 of all input variables were tested for each regression model. The behavior of RMSE and FS was evaluated and, for all the time horizons forecasting as well as for the raw/normalized variables, the behavior was similar. In this sense, the 4^{th} degree polynomial approximation was chosen, associating the necessary nonlinear sophistication with the computational load, in addition to relying on the model's ability to select variables, avoiding overfitting due to the high degree of the polynomial.

Figure 5 illustrates the variable selection capability of the LASSO where, for raw variables, only 46 out of the primary 165 (5 fourth order coefficients times 33 original) were selected. Variables that were already important in the other studied methods were kept, such as *zenith* and *minute*, associated with low order terms. For the less important variables, the coefficients of the terms of up to third-order outweighed the fourth-order terms, what tends to avoid overfitting. For the case of normalized variables, the selection was even more incisive, leaving only 28 predictors out of 165. The *zenith* lost its importance significantly, which is the main goal of the normalization procedure. The coefficients of up to third-order terms dominated again, totaling 18 out of the 28.

Ten-fold cross validation was applied through the *cv.glmnet* function. Grid search was applied to tune *lambda*, ranging from 10^{-2} until 10^{10}, with the value of the exponent varying uniformly in 100 steps. The best (lowest RMSE) λ obtained in the training set was then applied to the test set.

The results of nowcasting for the LASSO may be commented here. Like the previously presented methods, the overestimation occurred again, but slightly stronger than most of the other models. A subtle difference can be seen in the

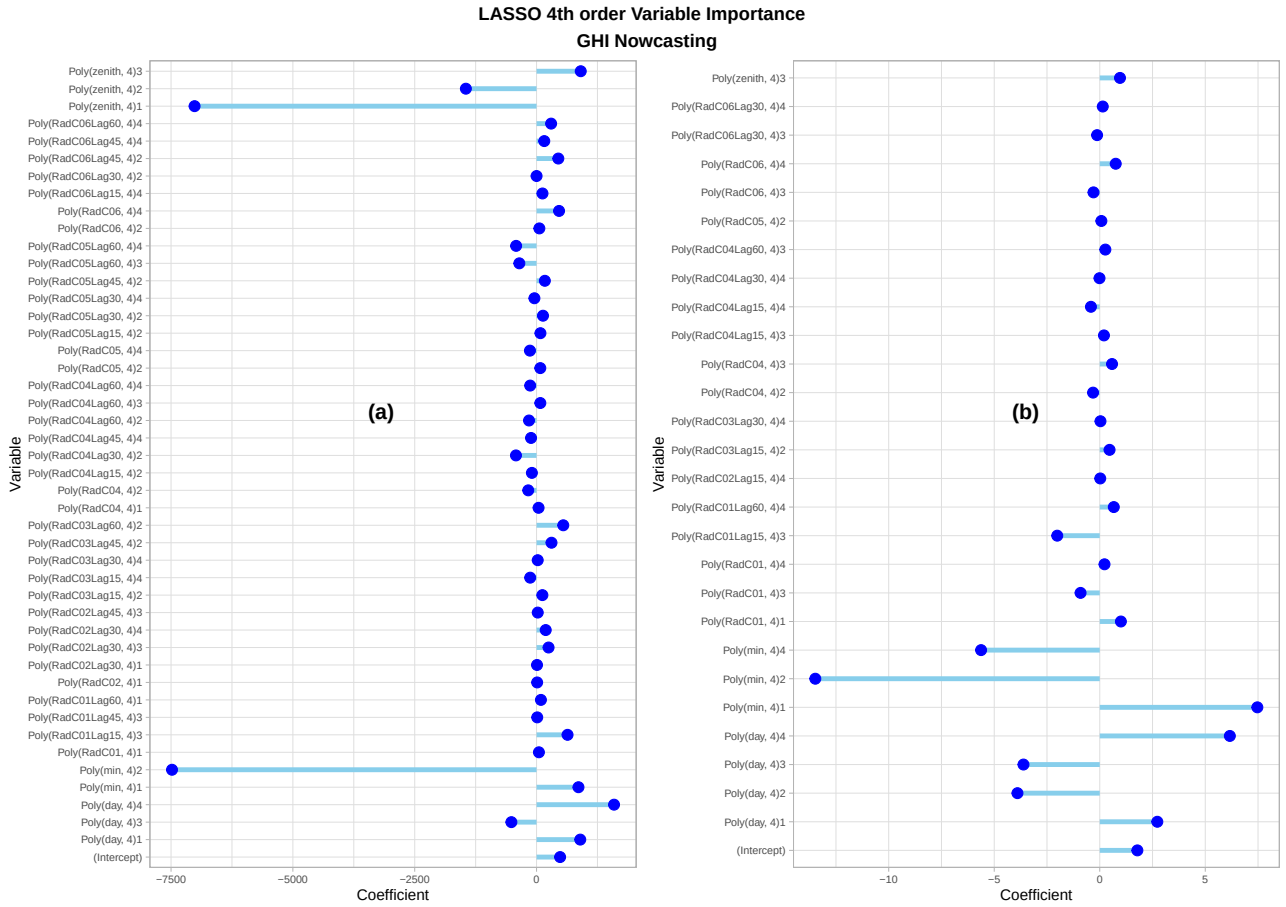

Figure 5. Variable importance for the LASSO model, applied in the GHI nowcasting. It is noteworthy the variable selection capability of the model, by the comparison of plot (a) (raw variables, already presenting a significant reduction in the variable quantity) and plot (b) (normalized variables case, which has left only 28 out of the original 165).

case of normalized variables, where predicted values greater than 750 W.m^{-2} may be seen.

3.5 Extreme Gradient Boosting (XGBoost)

The last model shown in this research, the eXtreme Gradient Boosting (XGBoost) presents itself as an extreme development of the ML approach in tree-based methods. It uses several techniques to improve the performance of trees, from the algorithm itself (tree pruning/ensemble) to training (additive training/learning rate). Its structure is inherently robust, which directly impacted the preprocessing phase. In this first step, the default settings of the hyperparameters of *caret* were used, with *center*, *scale*, *pca* and *spatial sign* preprocessing used separately or in combination.

The fine adjustment of the hyperparameters occurred right afterwards. First, a result was generated with the *caret* automatic grid search. The obtained hyperparameter values were taken and, one by one, they were progressively adjusted. For each hyperparameter, a search grid was tested looking for its optimum value. If this value lied on any extreme of the grid, its limits were then adjusted and the test was performed again, *ceteris paribus*. After that, the optimum value was retained and the search procedure was moved on to the next parameter. The parameters were tested in the following order: *nrounds*, *eta*, *max_depth*, *gamma*, *colsample_bytree*, *min_child_weight*, *subsample*.

Figure 6 presents the results of 15 minutes forecasting. it promptly makes clear the superiority of the results in comparison with the other studied methods. The points are more evenly distributed above and below the regression line, both for raw variables as well as the normalized ones. The tendency to overestimation has virtually disappeared. Finally, a discussion about the possible cause of this phenomenon is appropriate. Since the signal read by the satellite consists of the reflected radiation, both by clouds and/or by the soil (albedo), the algorithms can be confused by a given irradiance value as high, because when under clear sky or overcast conditions, a lot of solar radiation is reflected, in both cases arriving to the ABI. The XGBoost was able to detect the difference effectively, which significantly reduced the RMSE in the test set. This is reaffirmed by reading the generated histograms, which are very similar to those of the original data, both for raw variables and normalized ones. Discrepancies occur by the subtle overestimation for low GHI values and underestimation for high GHI values, noticeable by the observation of the regression line. Even so, values around and slightly above 1000

W.m^{-2} are perceived, which practically did not occur with the other methods.

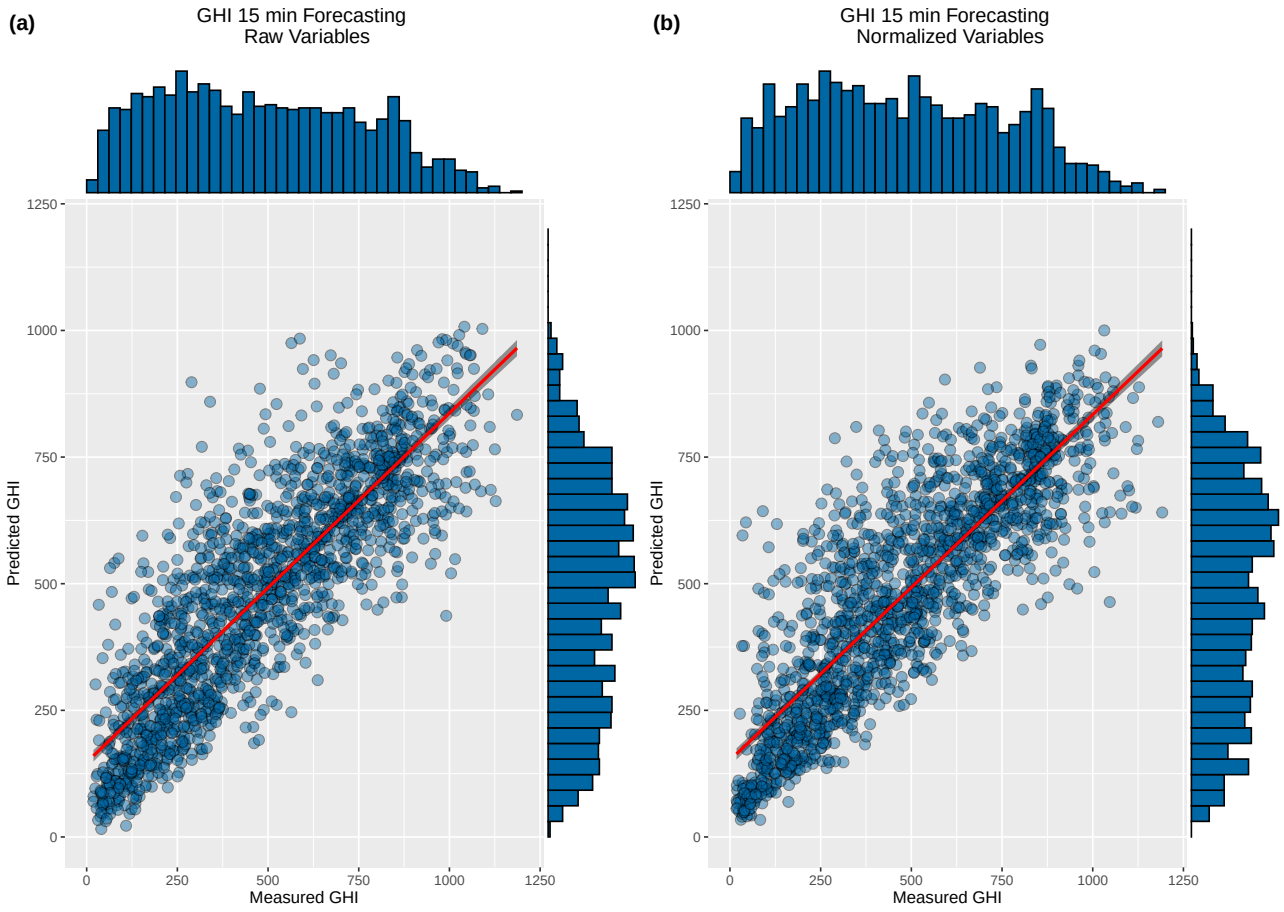

Figure 6. Measured *versus* predicted results for the GHI 15 minutes forecasting using XGBoost. The evenly points distribution around the regression line is noticeable. The histograms are more similar, for both the raw (a) and the normalized variables (b) cases.

Following the same approach as the previous ones, we evaluated the predictors regarding their importance. It can be seen the effect of normalization, which removes the higher weight of the *zenith* for the simulation with raw variables (a), giving more importance to the information of the 6 satellite channels, in the normalized variables case (b). It is also important to mention how channel 4 came to be relevant, a fact that had only happened in the ANN and SVM methods, accompanied by all of its time lags. From the normalized signals for the present time, the ones that showed the least impact were those of channels 2 and 5.

3.6 Overall comparison of the models performance

Since each applied model was individually presented and its results commented, a comparative assessment is necessary. In general, primarily in relation to the computational load of each method, most of them did not present processing times that can be considered prohibitive, even for slightly more advanced domestic platforms. The *k*-NN was noticeably faster, in addition to having a very simple parameterization, coming right after LASSO, certainly due to its linear nature already quite mature. With a medium load, both SVM and XGBoost showed themselves to have accessible use, for the case studied. The ANN, on the other hand, presented a significant computational load in comparison to the other options studied, with no associated gain in the now- forecasting performance, and should be avoided for this type of data structure, which has already been mentioned in the literature (Pazikadin *et al.*, 2020).

Regarding the now- and forecasting performance of the analyzed methods, Fig. 7 shows the plots of the RMSE (left column, (a) and (c)) and FS (right column, (b) and (d)) values. The results are also organized in relation to raw variables approach (top line, (a) and (b)) and normalized variables approach (bottom line, (c) and (d)). Taking into account what was presented at the end of the last paragraph, ANN did not have its computational load compensated by a better performance, with the worst result presented by all models, of 202.05 W.m^{-2}, for the nowcasting under the raw variables approach. Its best result was 182.87 W.m^{-2}, for nowcasting with normalized variables.

With slightly better performance and quite similar to each other, the other three methods appear, namely the SVM,

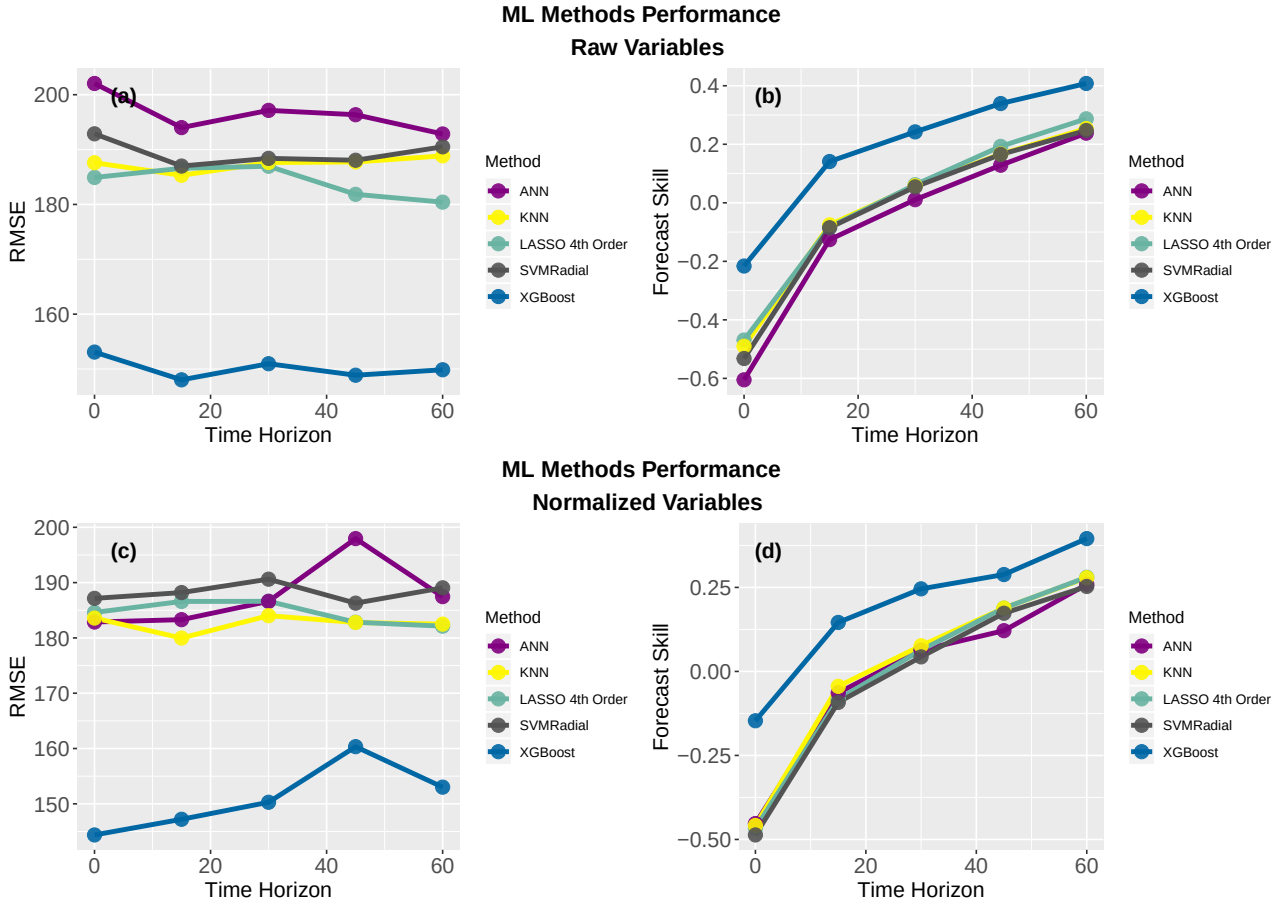

Figure 7. Overall comparison of the tested methods. RMSE lies in the left column plots (a) and (c), while FS is presented in the right column plots (b) and (d). The plots in the top row, (a) and (b), regard the raw variables approach, and in the bottom row, plots (c) and (d), the results of the normalized variables approach are shown.

k-NN and the LASSO. Their results were more consistent and regular, even though they were slightly worse than ANN in some cases with normalized variables. The best RMSE result was achieved by *k*-NN, 179.97 $W.m^{-2}$, for the 15 minutes forecasting using normalized variables. The worst result was obtained by SVM, with a RMSE of 192.87 W.m^{-2} for the normalized variables nowcasting.

With a performance significantly superior to the others in all cases studied, XGBoost achieved RMSE results of 144.37 W.m^{-2} for the nowcasting of the normalized variables case, 19.78 % better than the best case of the other methods. Although in this case the result was better than that of the raw variables, in all other forecasting results, i.e., in all time horizons, the results were numerically similar.

It is worth to mention the 45 minutes forecasting case of the normalized variables, where both ANN and XGBoost presented an anomaly of an out-of tendency RMSE increase. Most probably, for this time period, some slight disturbance was present in the training set, which was treated by these two most non-linear methods, which presented a very similar output. This did not happen to the other methods (*k*-NN, SVM and LASSO), which rely more in general geometric/algebraic relations. That fact impacted when the model was applied to the test set. Nevertheless, it is important to say that this does not represent any important variation on the results that could change the conclusions taken.

An analysis of the FS behavior can be done too. All the methods performed in a similar manner, with the normalized variables approach obtaining a better (for nowcasting) or equivalent (forecasting) result. All the values were negative for the nowcasting and 15 minutes forecasting, which implies that the dull persistence model would be a more desirable solution. But one must remember though, that for the dull persistence to work, a solarimetric station must be available, what definitely does not happen to the sattelite signal. As higher the time horizon is, dull persistence degenerates its performance, making the satellite signal modeling an even better option. Again, the XGBoost outperformed the other ones, being the only that presented a positive FS for the 15-min forecasting. Even so, for the nowcasting, the FS was -14.68 %, related to a RMSE of 144.37 W.m^{-2}, what is a competitive value even when compared to sky images approach, where values in the order of 143 W.m^{-2} are found in recent works, for 1 minute ahead forecasting, worsening the results when it comes until 5 minutes forecasting (Kamadinata *et al.*, 2019; Alonso-Montesinos *et al.*, 2015). For the FS, values around 25 % can be found for 1 hour forecasting (Feng and Zhang, 2020), which is lower than what we found for the

same time horizon (39.53 %).

4. Conclusion

This paper presented the application of five ML models in now- and forecasting GHI values, where the time horizons were 15, 30, 45 and 60 minutes ahead. The data set consisted of the pixel-wise 6 shortwave channels irradiance signals acquired through the GOES-16 images, their previous temporal values, the zenith angle, the day of the year and the minute of the day, suming 33 predictors. From this study, the following conclusions can be drawn:

- Four of the applied methods, namely ANN, *k*-NN, the LASSO and SVM, have shown the general trend to overpredict the values for all the time horizons studied, including the nowcasting. Only the XGBoost could overcome this flaw, thus presenting a significant RMSE decrease, naturally accompanied by the highest FS values. Specifically, the predicted GHI values were higher for the lower ranges and lower for the higher ranges;

- The variable normalization approach, in general, improved the results, decreasing the RMSE values. After this detrending procedure, the variable importance evaluation demonstrated a gain in the satellite signals relevance for the models. For the XGBoost this conclusion cannot be stated though, since in this case the error values increased over the time horizon;

- Some more variable preprocessing may be necessary to improve results. The *scale*, center, *pca* and *spatial sign* were tested, and implemented when any of them improved the results. No overall conclusion can be drawn about it, since this approach showed itself model dependent;

- All the methods presented an increase in FS over the time horizons, being the 60 minutes ahead the case that the higher values were achieved. All the methods performed worse than persistence for the nowcasting case, presenting negative values for the FS. Only the XGBoost was better than persistence in the 15 minutes case, and all of them performed better in the other time horizons.

5. ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and accomplished with the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq) - Grant No. 305456/2019-9, both Brazilian governmental agencies.

6. REFERENCES

, 2020. URL http://sonda.ccst.inpe.br/basedados/petrolina.html.

, 2020. "Renewable Energy Statistics 2019". URL /publications/2019/Jul/Renewable-energy-statistics-2019.

Alonso-Montesinos, J., Batlles, F. and Portillo, C., 2015. "Solar irradiance forecasting at one-minute intervals for different sky conditions using sky camera images". *Energy Conversion and Management*, Vol. 105, pp. 1166 – 1177.

Ayodele, T., Ogunjuyigbe, A., Amedu, A. and Munda, J., 2019. "Prediction of global solar irradiation using hybridized k-means and support vector regression algorithms". *Renewable Energy Focus*, Vol. 29, pp. 78 – 93.

Caldas, M. and Alonso-Suárez, R., 2019. "Very short-term solar irradiance forecast using all-sky imaging and real-time irradiance measurements". *Renewable Energy*, Vol. 143, pp. 1643 – 1658.

Chow, C.W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J. and Washom, B., 2011. "Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed". *Solar Energy*, Vol. 85, No. 11, pp. 2881 – 2893.

Cornejo-Bueno, L., Casanova-Mateo, C., Sanz-Justo, J. and Salcedo-Sanz, S., 2019. "Machine learning regressors for solar radiation estimation from satellite data". *Solar Energy*, Vol. 183, pp. 768 – 775.

de Abreu [Melo Junior], F.E., [de Moura], E.P., Rocha], P.A.C. and [de Andrade], C.F., 2019. "Unbalance evaluation of a scaled wind turbine under different rotational regimes via detrended fluctuation analysis of vibration signals combined with pattern recognition techniques". *Energy*, Vol. 171, pp. 556 – 565.

de Minas e Energia, M., 2020. "2019 - ministério de minas e energia". URL http://www.mme.gov.br/web/guest/secretarias/energia-eletrica/publicacoes/boletim-de-monitoramento-do-

Demolli, H., Dokuz, A.S., Ecemis, A. and Gokcek, M., 2019. "Wind power forecasting based on daily wind speed data using machine learning algorithms". *Energy Conversion and Management*, Vol. 198, p. 111823.

Elmaz, F., Özgün Yücel and Mutlu, A.Y., 2020. "Predictive modeling of biomass gasification with machine learning-based regression methods". *Energy*, Vol. 191, p. 116541.

Fan, J., Wang, X., Zhang, F., Ma, X. and Wu, L., 2020. "Predicting daily diffuse horizontal solar radiation in various climatic regions of china using support vector machine and tree-based soft computing models with local and extrinsic climatic data". *Journal of Cleaner Production*, Vol. 248, p. 119264.

Feng, C. and Zhang, J., 2020. "Solarnet: A sky image-based deep convolutional neural network for intra-hour solar forecasting". *Solar Energy*, Vol. 204, pp. 71 – 78. ISSN 0038-092X.

Google, 2020. "Google cloud sdk". URL `https://cloud.google.com/sdk?hl=pt-br`.

Hotelling, H., 1936. "Relations between two sets of variates". *Biometrika*, Vol. 28, No. 3/4, p. 321. ISSN 00063444.

Inman, R.H., Pedro, H.T. and Coimbra, C.F., 2013. "Solar forecasting methods for renewable energy integration". *Progress in Energy and Combustion Science*, Vol. 39, No. 6, pp. 535 – 576.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer.

Kallio-Myers, V., Riihelä, A., Lahtinen, P. and Lindfors, A., 2020. "Global horizontal irradiance forecast for finland based on geostationary weather satellite data". *Solar Energy*, Vol. 198, pp. 68 – 80.

Kalluri, S., Alcala, C., Carr, J., Griffith, P., Lebair, W., Lindsey, D., Race, R., Wu, X. and Zierk, S., 2018. "From photons to pixels: Processing data from the advanced baseline imager". *Remote Sensing*, Vol. 10, No. 2, p. 177. ISSN 2072-4292.

Kamadinata, J.O., Ken, T.L. and Suwa, T., 2019. "Sky image-based solar irradiance prediction methodologies using artificial neural networks". *Renewable Energy*, Vol. 134, pp. 837 – 845.

Kuhn, M., 2008. "Building predictive models in r using the caret package". *Journal of Statistical Software*, Vol. 28, No. 1, p. 1–26. ISSN 1548-7660.

Larson, D.P., Li, M. and Coimbra, C.F.M., 2020. "Scope: Spectral cloud optical property estimation using real-time goes-r longwave imagery". *Journal of Renewable and Sustainable Energy*, Vol. 12, No. 2, p. 026501.

Lima, M.A.F., Carvalho, P.C., Fernández-Ramírez, L.M. and Braga, A.P., 2020. "Improving solar forecasting using deep learning and portfolio theory integration". *Energy*, Vol. 195, p. 117016.

NASA, N.., 2020. "Mission overview goes-r series". URL `https://www.goes-r.gov/mission/mission.html`.

Pazikadin, A.R., Rifai, D., Ali, K., Malik, M.Z., Abdalla, A.N. and Faraj, M.A., 2020. "Solar irradiance measurement instrumentation and power solar generation forecasting based on artificial neural networks (ann): A review of five years research trend". *Science of The Total Environment*, Vol. 715, p. 136848.

Pearson, K., 1901. "Liii. on lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Vol. 2, No. 11, p. 559–572. ISSN 1941-5982, 1941-5990.

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

Platform, G.C., 2020. "Big query - google cloud platform". URL `https://console.cloud.google.com/marketplace/details/n`

Rocha, P.A.C., Fernandes, J.L., Modolo, A.B., Lima, R.J.P., da Silva, M.E.V. and Bezerra, C.A.D., 2019. "Estimation of daily, weekly and monthly global solar radiation using anns and a long data set: a case study of fortaleza, in brazilian northeast region". *International Journal of Energy and Environmental Engineering*, Vol. 10, No. 3, pp. 319–334.

Saidi, K. and Omri, A., 2020. "The impact of renewable energy on carbon emissions and economic growth in 15 major renewable energy-consuming countries". *Environmental Research*, Vol. 186, p. 109567.

Serneels, S., De Nolf, E. and Van Espen, P.J., 2006. "Spatial sign preprocessing: a simple way to impart moderate robustness to multivariate estimators". *Journal of Chemical Information and Modeling*, Vol. 46, No. 3, pp. 1402–1409.

Tibshirani, R., 1996. "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288.

Wolpert, D.H. and Macready, W.G., 1997. "No free lunch theorems for optimization". *Trans. Evol. Comp*, Vol. 1, No. 1, p. 67–82.

Yang, D., 2020. "Choice of clear-sky model in solar forecasting". *Journal of Renewable and Sustainable Energy*, Vol. 12, No. 2, p. 026101. ISSN 1941-7012.

Yang, D. and Bright, J.M., 2020. "Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years". *Solar Energy*.

Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T. and Coimbra, C.F., 2018. "History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining". *Solar Energy*, Vol. 168, pp. 60 – 101. Advances in Solar Resource Assessment and Forecasting.

Yang, D., Ye, Z., Lim, L.H.I. and Dong, Z., 2015. "Very short term irradiance forecasting using the lasso". *Solar Energy*, Vol. 114, pp. 314 – 326.

Zagouras, A., Pedro, H.T. and Coimbra, C.F., 2015. "On the role of lagged exogenous variables and spatio–temporal correlations in improving the accuracy of solar forecasting methods". *Renewable Energy*, Vol. 78, pp. 203 – 218.

Zhao, X., Wei, H., Wang, H., Zhu, T. and Zhang, K., 2019. "3d-cnn-based feature extraction of ground-based cloud images for direct normal irradiance prediction". *Solar Energy*, Vol. 181, pp. 510 – 518.

7. RESPONSIBILITY NOTICE

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