2022 IEEE Multi-conference on Natural and Engineering Sciences for Sahel's Sustainable Development (MNE3SD) | 978-1-6654-2152-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/MNE3SD53781.2022.9723309

Prediction of daily direct solar energy based on XGBoost in Cameroon and key parameter impacts analysis

Douanla Alotse Yaulande
Physics
IMSP-UAC
Dangbo, Bénin
yaulande.douanla@imsp-uac.org

Dembélé André
Operations Research & Optimization
USTTB
Bamako, Mali
andredembele@ymail.com

Mamadou Ossénatou
Physics
IMSP-UAC
Dangbo, Bénin
ossenatou.mamadou@gmail.com

Lenouo André
Physics
Université de Douala
Douala, Cameroun
lenouo@yahoo.fr

***Abstract*—This study explores the ability of Extreme Gradient Boosting (XGBoost) to predict the direct normal irradiation (DNI) under clear sky conditions in Cameroon. The satellite data used are DNI clear sky, air Temperature, Relative Humidity, Wind Speed, Wind direction, irradiation at Top of Atmosphere (TOA) and Aerosol Optical Depth at 550 nm (AOD550) for each aerosol type (Black Carbon : BCAOD550; Organic Matter : OMAOD550; Sea Salt : SSAOD550; Sulphate : SUAOD550 and Dust: DUAOD550). To achieve this aims and build a worst case prediction scenario, K-means clustering algorithm with Elbow and Silhouette analysis are used to select training and validation data sets. The coefficient of determination R^2, root mean square error $RMSE$ and the interpretation of the model outputs in the light of the state of the art confirm the robustness of the used model.**

The interpretation of the XGBoost outputs using the Shapley's value shows that the amount of energy in the study area is most impacted by DUAOD550, OMAOD550, temperature, SUAOD550, TOA, SSAOD550 and relative humidity respectively. Results suggest also that, if dust and organic matter aerosols are present in the same proportion, the attenuation produced by them can be 4 to 10 times higher than those induced by black carbon and sea salt aerosols.

***Index Terms*—XGBoost, Irradiation, Aerosols, Shapley, Shap, Cameroon.**

The first author is grateful to *German Academic Exchange Service* (DAAD) for the PhD fellowship (section ST32, No 91722008). This work was carried out with the aid of a grant from UNESCO and the International Development Research Centre, Ottawa, Canada. The views expressed herein do not necessarily represent those of UNESCO, IRDC or its Board of Governors. Authors also thank to SoDa, MERRA and CAMS database administrators for providing open access to their databases. Finally, authors are grateful to the "Institut de Mathématiques et de Sciences Physiques (IMSP)", the African Centre for Excellence in Mathematic Sciences, Informatics and Applications (ACE-MSIA) of the University of Abomey-Calavi (Bénin).

I. Introduction

Solar radiation is the main and most interesting source of energy on Earth because it is naturally available (free, abundant), clean, inexhaustible [1], [2]. It has also a great impact on the climate [3]. However, the amount of incident solar radiation received is attenuated in the atmosphere by several atmospheric components [1]. The process of solar energy attenuation by atmospheric particles varies from site to site depending on the natural and anthropogenic loads on the atmosphere. Among the atmospheric particles mentioned here are the atmospheric aerosols generally characterized by their optical depth (AOD) which measures the attenuation of direct radiative effect by these particles. AOD depends greatly on the study area but also on aerosol's position relative to the sources, seasonal variations as well as climatic conditions [4], [5]. Indeed, atmospheric aerosols are among the most dominant solar radiation attenuating factors [6] and many studies have addressed this [7], [8], [9]. In order to ensure long-term investments in solar energy technologies, several works have been carried out to predict the availability of solar energy [10] using either empirical or smart machine learning models [11], [1].

In this idea, we propose in this work to investigate the ability of the Extreme Gradient Boosting (XGBoost) model to forecast the daily direct solar energy over Cameroon under clear sky conditions as a function of different types of aerosols and other meteorological variables. XGBoost can give better performance (accuracy and runtime) than Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Random Forest (RF) [11]. In a recent work done by [1], several machine learning models were built among

which one of the best for predicting monthly solar radiation was the XGBoost model. In addition, [12] evaluated the performance of three machine learning models to predict daily diffuse solar radiation. These authors concluded that XGBoost model stands out for its strong performance and stability.

The rest of the paper is structured as follows: Section II describes the study area, data and clustering approach used to address the main stated goal. In section III, we explain machine learning method used in this study. This includes the XGBoost algorithm and Shapley values useful to explain the relative contribution of each parameter to the prediction. Section IV provides some hyperparameter best values for the model, builds the model and validates it. Finally, section V provides the results and we conclude in section VI.

II. Study Area and Data

The study area is Cameroon (Figure 2), country of Central part of Africa with an area of about 475.000 km^2 and has the highest mountain ranges in Africa. The country is bordered by the Congo basin, Lake Chad and the Atlantic Ocean and divided into two main climatic regions: the equatorial and sub-equatorial regions (in the south) and the tropical regions (in the north).

Data were downloaded from three different sources (Table 1.) for the period 2005-2019.

The Copernicus Atmosphere Monitoring Service radiation service (CAMS Radiation) data cover the period from 2004-02-01 up to 2 days ago. As with CAMS Radiation, CAMS-OAD provided data of aerosol optical depth at 550 and 1240 nm wavelengths and those of different aerosol types, such as Black Carbon aerosols, Dust, Organic Carbon (Organic Matter), Sea Salt, Sulfate at 550 nm. The Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2), provides since January 1980 time series of Wind speed and direction (at 10 m), and finally both air Temperature and Relative Humidity (at 2 m).

The Angstrom exponent (alpha) and Angstrom's turbidity coefficients at 550 nm (beta550) were calculated using the AOD value between 550 and 1024 nm. Direct (or Beam) Normal Irradiation (BNI.Clear.Sky) is the target variable, and the remaining variables including alpha and beta550 represent the features (inputs) as evidenced by the pearson correlation matrix (see Figure 1).

This study covers the period from January 1, 2015, to December 31, 2019. From latitude 1.6° to 13.1° and longitude 8.38° to 18.38°, with a 50km step size, we obtain 504 observation sites where only 173 are entirely enclosed in the study area and provide our sample data.

TABLE I: Data sources and variables used.

SOURCE	DATA (2005-2019)
MERRA-2	air Temperature, Relative Humidity Wind Speed and Wind direction
CAMS-AOD	BCAOD550, DUAOD550 OMAOD550, SSAOD550 SUAOD550, AOD550 and AOD1024
CAMS Radiation	BNI.Clear.Sky, TOA

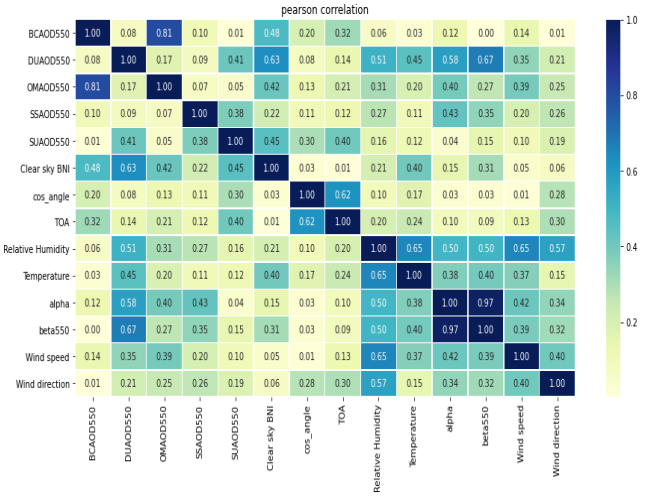


Fig. 1: Pearson correlation matrix between meteorological data and other variables used in this study.

III. Methods

A. Extreme Gradient Boosting (XGBoost)

Classical boosting algorithms can cause overfitting. To remedy this, a regularised gradient boosting algorithm is presented by [13]. XGBoost algorithm is an aggregates trees algorithm. At each iteration, the new tree receives the error made by the previous tree. Thus, even if individual trees have low predictive accuracy, the decision rule built by summing the results of each tree is very reliable.

The principle of XGBoost algorithm is as follows [13], [14], [15], [16]. For a given dataset B with n features and p explanatory variables $B = (x_i, y_i)_{1<i<n}$, where $|B| = n, x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$. The output prediction model is defined as:

$$\hat{y}_i = \theta(x_i) = \sum_{k=1}^{M} f_k(x_i), f_k \in G \tag{1}$$

where G is the regression trees space and $|G| = M$. Thus, with a regression tree $f \in G$ and a sample x, $f(x)$ is the

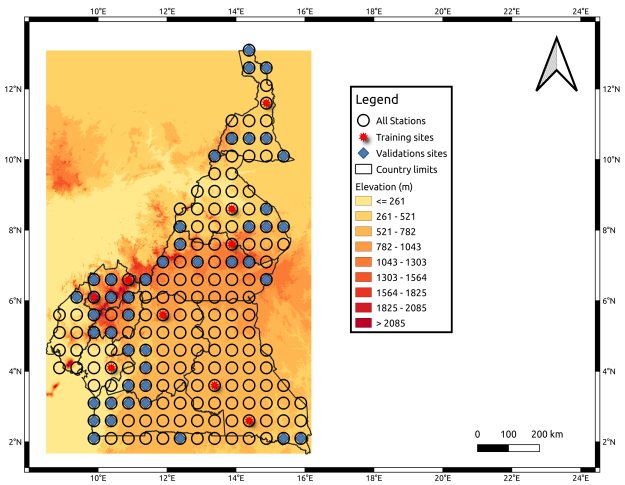


Fig. 2: Topography of the study area and selected sites according to K-means algorithm with Elbow and Silhouette analysis. Red color marks the centroids where training set is extracted and blue color shows the sites where validation data are extracted.

score w_{if} assigned to the leaf (indexed by i^f) to which x belongs.

XGBoost uses Newton boosting which defines optimal parameters by minimizing loss function (θ) given by

$$T(\hat{y}_i) = \sum_{i=1}^{n} l(\hat{y}_l, y_i) + \sum_{k=1}^{M} \tau(f_k), \tag{2}$$

$$\tau(f_k) = \gamma N + \frac{1}{2}\alpha ||w_{kf}||^2, \tag{3}$$

where γ and α are model's penalty parameters. This model includes functions as parameters and therefore cannot be optimized in the classical way. Thus, the model is changed for another model that is trained in an additive manner.

At a given iteration s, f_s is added to minimize the loss function:

$$T^s = \sum_{i=1}^{n} \left[l\big(y_i, \hat{y}_i^{s-1} + f_s(x_i)\big)\right] + \tau(f_s), \tag{4}$$

where $\hat{y}_l^{(s)}$ stands for the prediction at iteration s of the l-th instance.
A second order Taylor polynomial of T^s gives

$$T^s = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{s-1}) + g_i f_s(x_i) + \frac{1}{2} h_i f_s^2(x_i)\right] + \tau(f_s) \tag{5}$$

where $g_i = \partial_{\hat{y}_i^{s-1}} l(y_i, \hat{y}_i^{s-1})$ and $h_i = \partial^2_{\hat{y}_i^{s-1}} l(y_i, \hat{y}_i^{s-1})$.

Now replacing (3) in (4), we can rewrite (5) as

$$\tilde{T}^{(s)} = \sum_{i=1}^{n} [g_i f_s(x_i)) + \frac{1}{2} h_i f_s^2(x_i)] + \gamma N + \frac{1}{2}\alpha \sum_{k=1}^{M} w_j^2 \tag{6}$$

$$= \sum_{k=1}^{M} \Big[(\sum_{i \in C_k} g_i) w_k + \frac{1}{2} (\sum_{i \in C_j} h_i + a) w_k^2 \Big] + \gamma N \tag{7}$$

where

$$C_k = \{ i \mid r(x_i) = k, \ i = 1, \ldots, n \} \tag{8}$$

is the set of samples that belong to the same leaf k of a certain regression tree f_r those structure is given by r (r assigns to each sample the leaf that corresponds to it in the tree f_r).
For a given structure r, the optimal score w_k^* of a leaf k is

$$w_k^* = -\frac{\sum_{i \in C_k} g_i}{\sum_{i \in C_k} h_i + \alpha}. \tag{9}$$

Finally, the optimal objective function value at step s is [13]:

$$\tilde{T}^{(s)}(r) = -\frac{1}{2} \sum_{k=1}^{M} \frac{(\sum_{i \in C_k} g_i)^2}{\sum_{i \in C_k} h_i + \alpha} + \gamma N, \tag{10}$$

a function that measures the impurity of the tree r.
Since one can measure the quality of a tree, how could one improve this value at each step? A greedy algorithm allows, from each leaf, to choose which branching (or disaggregation) to do. Suppose that leaf C_k (see (8)) is split into C_k^l and C_k^r i.e. $C_k = C_k^l \cup C_k^r$. Then the loss reduction or Gain after the split is [13]:

$$T_{split} = \frac{1}{2} \Big[\frac{(\sum_{i \in C_k^l} g_i)^2}{\sum_{i \in C_k^l} h_i + \alpha} + \frac{(\sum_{i \in C_k^r} g_i)^2}{\sum_{i \in C_k^r} h_i + \alpha} - \frac{(\sum_{i \in C_k} g_i)^2}{\sum_{i \in C_k} h_i + \alpha} \Big] - \gamma. \tag{11}$$

We use the following parameters within our XGBoost regressor (they also provide a stopping condition for the algorithm):

- `learning_rate` (learning rate) : Step size shrinkage used in update to prevents overfitting;
- `max_depth` (maximum depth of a tree);
- `gamma` (γ): Minimum loss reduction required to make a further partition on a leaf node of the tree.

Their optimal values are provided after a bayesian optimization [17]. We choose the values for some hyperparameters (`subsamble=` 0.8 (subsample ratio of the training instances), `learning_rate=eta=` 0.1, `eval_metric=rmse`) and the other use their default settings [18].

B. SHAP values

Whether in regression or in classification, it is often very difficult to explain efficiently the outputs model (machine learning results), i.e. to explain the importance or contribution of each parameter. To remedy this, there are some techniques to better explaining output models like Lime, DeepLift and Layer-Wise Relevance Propagation for deep learning, Classic Shapley Value Estimation and SHAP (SHapley Additive exPlanations) (see [19]). SHAP unifies these predecessor methods and show improved computational performance and better consistency and accuracy than previous approaches [19]. SHAP is based on the Shapley values which determines the individual contribution of player in a cooperative game (collaborative team). Explaining the prediction using Shapley values involves assigning a coefficient to each input variable indicating how each contributed to the shift in the prediction. So, SHAP interprets a predicted value as the sum of each variable contributions ϕ_k plus the base value ϕ_0:

$$y_{pred} = \phi_0 + \sum_{k=1}^{M} \phi_k z_k^{'}. \tag{12}$$

With y_{pred} the model's predicted value, ϕ_0 the base model's model (defined as the average of all predictions in the dataset), $z^{'} \epsilon \{0,1\}^M$ indicating whether the variable is observed $z_k^{'} = 1$ or unknown $z_k^{'} = 0$ (see [19]). Let N stands for the number of features of set E put together to obtain a result $v(E) \in \mathbb{R}$ and s the number of features (other than k) of a subset F (containing k) within E. Knowing the contribution $v(F)$ of each subset $F \subseteq E$, the marginal value of k in each subset F is $[v(F) - v(F \setminus \{k\}]$ and the shapley value ϕ_k of k is expressed in [20] as:

$$\phi_k = \frac{1}{N!} \sum_{F \subseteq N \setminus \{k\}} |N|!(|N| - |s| - 1)![v(F) - v(F \setminus \{k\})] \tag{13}$$

Shap algorithm uses in this work comes from `https://github.com/slundberg/shap.git`.

IV. Implementation of the Forecasting Model

A. Data preparation

Two objectives guide the choice of the training and validation data: firstly selecting among the 173 data points the most representative of the population; and secondly preparing the evaluation data set that would describes a worst-case forecast scenario (the farthest from the training data).
Thus, with the initial data from the 173 sites (with a time step of 3 hours), we generate the quarterly means data. Kmeans algorithm [21], [22] allows to perform classifications for a number of clusters k ranging from 4 to 20. **Elbow** and **Silhouette** methods are used to select the optimal value ($k = 9$) as the optimal number of clusters. The training data is thus the set of data from the 9 sites closest to the cluster centers obtained. Concerning the data used for the validation, they correspond to the set of 5 sites per cluster farthest from the centers.

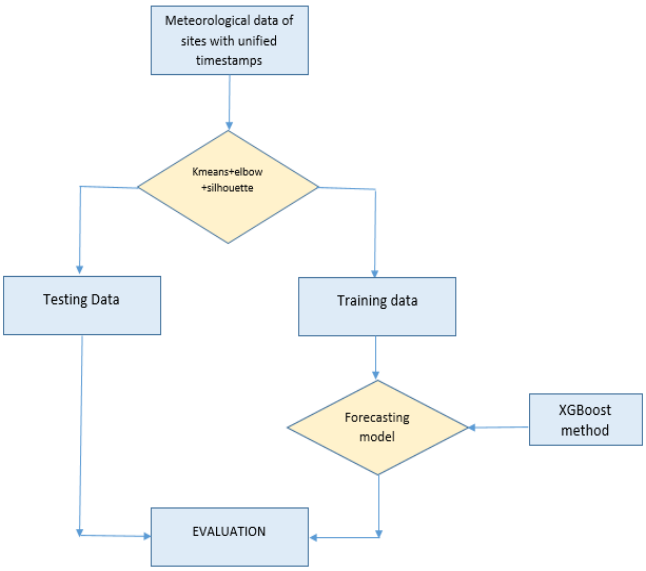


Fig. 3: Methodology diagram adopted in the study.

B. XGBoost setting : Hyperparameters optimization

Several methods allow to find hyperparameter's values for which the prediction models give better results. Among those methods are: the grid search method, random search and bayesian optimization. Bayesian optimization is very popular. It tries to evaluate the objective function as less possible in order to reduce its computational burden. Bayesian optimization algorithm is provided by `bayes_opt` (a python official package [17]. The number of iterations is 100 (`n_iter`) with 8 steps of random exploration (`init_points`).

TABLE II: XGBoost settings

$N°$	Parameter	Search space	best value
1	learnig_rate	$[0, 1]$	0.14
2	max_depth	$[2, 20]$	10
3	gamma	$[0, 1]$	0.003
4	n_estimators	$[700, 1400]$	1103

V. Results

After the identification of the validation and training dataset, we generate the daily average dataset that will be

used for the rest of the study. So we have $(9 * 5478)$ points for the training and $(45 * 5478)$ points for the validation. The training data represents $1/10$ of the data used and the validation data $9/10$. With that setting and those obtained after bayesian optimization, XGBoost provides $R^2 = 0.84$ and $RMSE = 0.55$ and allows to start the interpreting the predictions of model.

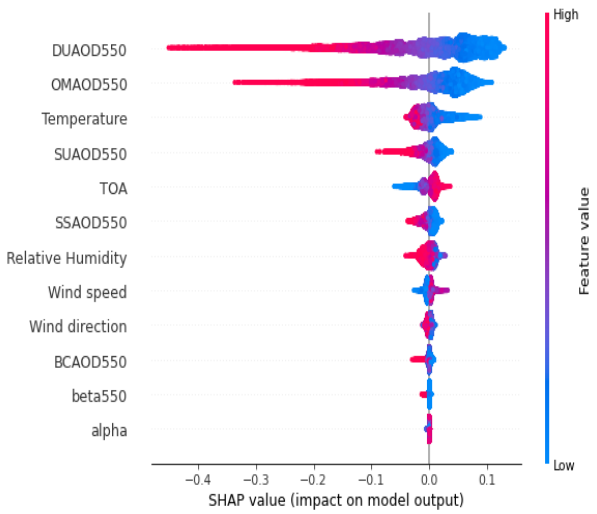


Fig. 4: Summary plot

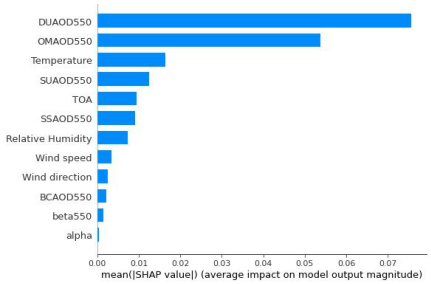
Figure 4. is a summary plot provided by training dataset which is about 49302 points. Each point of the cloud corresponds to a point in the training dataset. Features are ranked in descending order importance. The horizontal axis indicates the effect of that variable on the prediction. Color shows whether variable value of that point is high (in red) or low (in blue).

For a given feature, Figure 4 highlights the type of correlation with the target variable according to the scatter of the point cloud. A horizontal dispersion for an given feature denotes a linear correlation while a dispersion along the vertical line proves a more complex correlation. Figure 4 shows that high value of optical thicknesses of different types of aerosols in particular (DUAOD550, OMAOD550 and SUAOD550) and those of relative humidity makes the value of energy (direct normal irradiation in clear sky condition) low. While, high value of irradiation at the top of atmosphere (TOA) makes the value of the prediction high.The values of beta, alpha, BCAOD550 and Wind direction have little influence on the final model.

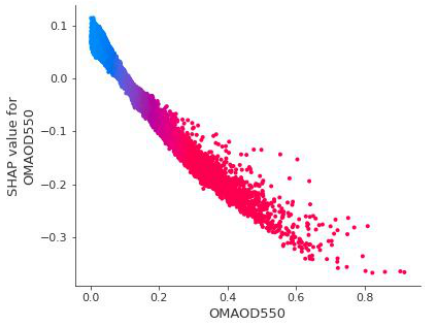
Figure 4. shows the global importance of each feature. This feature importance is expressed as the average of the absolute values of the shap values. The nature of correlation between two features and the impact of their interaction on the direct irradiation will be analyzed with the partial dependence plot (PDP or PD plot) (Figures 6-13). In these figures, the horizontal axis indicates the value from the training dataset and the vertical axis (on the right) shows the impact of that value on the prediction. PD plot can also highlights the impact of a feature on the prediction.

Figure 1. shows that aerosols (DUAOD550, OMAOD550 and SSAOD550) are the most important sources affecting direct energy. It also shows the major influence of air Temperature, TOA and Relative Humidity. DUAOD550 and OMAOD550 greatly attenuate the direct radiation (see the two first lines of Figure 1.). The pearson correlation matrix shows the level of absolute linear correlation between the variables. We can thus note a strong linear correlation between BNI.clear.sky and: DUAOD550 (0.63) followed by BCAOD550 (0.48), SUAOD550 (0.45), OMAOD550 (0.42), Temperature (0.40) and SSAOD550 (0.22) (Figure 1). The analysis of the XGBoost model with SHAP reveals that the most important variables are DUAOD550, OMAOD550, Temperature, SUAOD550. The results provided by SHAP are much more accurate because they take into account the effects of the interactions between the variables and not simply the effect of their individual influences added. Under clear sky conditions, the irradiation is mainly affected by aerosols. (Figure 4; is Temperature more important than SUAOD550 ? The answer is no.) Temperature is influenced by several other processes even in the absence of significant radiation such as evapotranspiration and other microclimates induced by other variables. A such relative impact can be seen through the pearson correlation matrix.

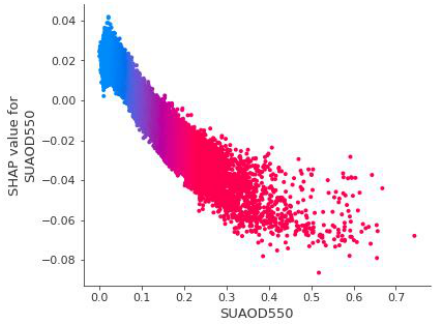
According to [23], the majority of climate models take into account only the absorbing effects of aerosols. Thus, in the same proportions, the attenuation produced by DUAOD550 and OMAOD550 can be 4 to 10 times higher than that produced by BCAOD550 and SSAOD500 (Figures 10, 6, 9, 8). The relationship between TOA and Relative humidity (positive correlation) has been showed in Figure 12; while one can clearly identify the relevant influence of TOA on relative humidity in (Figures 11 and 13).



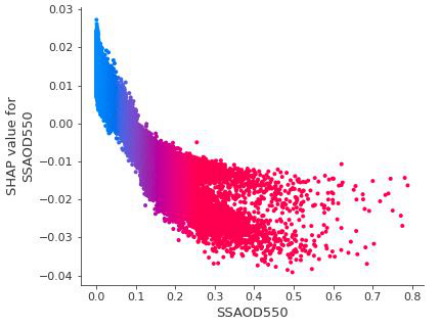
(a) Fig. 5: Global importance plot



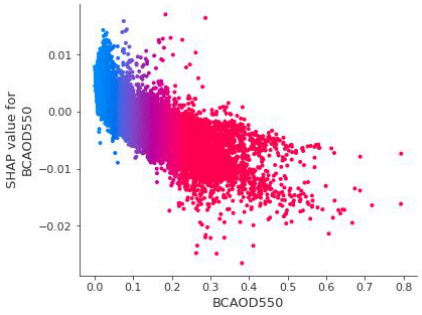
(b) Fig. 6: OMAOD550 importance



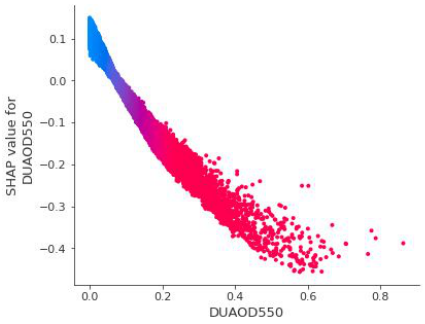
(c) Fig. 7: SUAOD550 importance



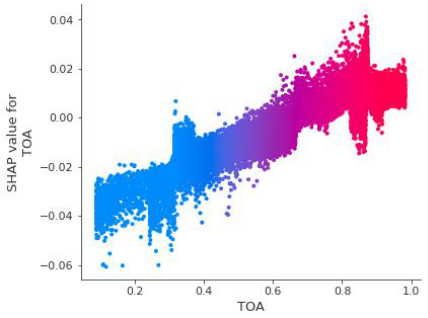
(d) Fig. 8: SSAOD550 importance



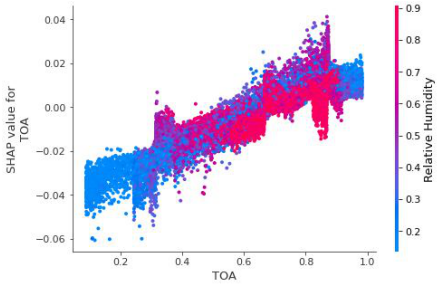
(e) Fig. 9: BCAOD550 importance



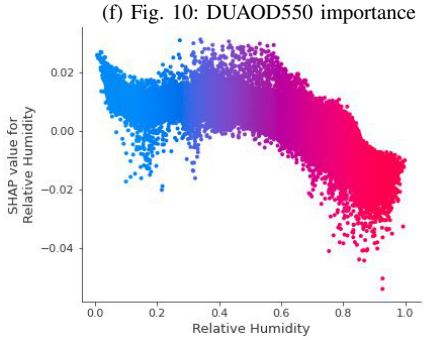
(f) Fig. 10: DUAOD550 importance



(g) Fig. 11: TOA importance



(h) Fig. 12:TOA & Relative Humidity PD plot



(i) Fig. 13: Relative Humidity importance

VI. Conclusion

This work shows that XGBoost is a robust tool for predicting direct energy at daily scale over Cameroon (Central Africa). The scenario used is to test the ability of the model by using a small training data size. We demonstrated that dealing with this smart choice of the training data, a good prediction of the daily direct normal solar energy can be obtained. Besides, we proved using the shap method that aerosols are the main sources of the attenuation of solar energy relatively to the meteorological variables used in the study area. It has been found that the attenuation produced by DUAOD550 and OMAOD550 can be 4 to 10 times higher than that produced by BCAOD550 and SSAOD500. Beyond the already known interactions, it can be noticed that even when TOA is positively correlated with direct energy, unlike relative humidity, the combined influence of both is positively correlated with direct solar energy.

Data access

This projet was possible thanks to free and open access data available following these links:

- http://www.soda-pro.com/web-services/meteo-data/merra
- http://www.soda-pro.com/web-services/radiation/cams-radiation-service
- http://www.soda-pro.com/web-services/atmosphere/cams-aod

REFERENCES

[1] H. Liexing, K. Junfeng, W. Mengxue, F. Lei, Z. Chunyan, Z. Zhaoliang, "Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events," Frontiers in Earth Science, vol. 9, pp. 202, 2021.

[2] J. Yue, S. Yongjun, W. Fengchun, L. Pengcheng, H. Shuyi, "Global solar radiation modeling using different machine learning and empirical models in Northeast China," 2021.

[3] B. Christian, R. Markus,T. Enrico, C. Philippe, J. Martin, C. Nuno, R. Christian, A. MAltaf, B. Dennis, B. Gordon, others, "Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate," Science, vol. 329, pp. 834–838, 2010.

[4] C. Cheng, O. Dubovik, G. Schuster, L. Gregory, D. Fuertes, Y. Meijer, J. Landgraf, Y. Karol and Z. Li. "Characterization of temporal and spatial variability of aerosols from ground-based climatology: towards evaluation of satellite mission requirements," Journal of Quantitative Spectroscopy and Radiative Transfer, vol. 268, pp. 107627, 2021, Elsevier.

[5] F. Thomas, D. Larry, G. Shannon, "The spatial and temporal variability of aerosol optical depths in the Mojave Desert of southern California," Remote sensing of environment, vol. 107, pp. 54–64, 2007.

[6] L. KJ, W.Feng, JC. Xu and PX. Gao, LH.Yang, HF. Liang, LS. Zhan, "Why is the solar constant not a constant?," The Astrophysical Journal, vol. 747,pp. 135, 2012.

[7] A. Mani, O. Chacko, "Attenuation of solar radiation in the atmosphere," Sun II, pp.2208–2212, 1979.

[8] G. Antonio, V. Eunice, E. Álvarez-Álvarez, G. Juan, X. Jorge, S. María , "Attenuation processes of solar radiation. Application to the quantification of direct and diffuse solar irradiances on horizontal surfaces in Mexico by means of an overall atmospheric transmittance,", Renewable and Sustainable Energy Reviews, pp.93–106, 2018.

[9] P. Jesús, A. Joaquin, L. Gabriel, B. Jesús, Bosch, L. Juan, B. Javier, C. Elena, F. Jesús, B. Francisco, "Modelling atmospheric attenuation at different AOD time-scales in yield performance of solar tower plants, " AIP Conference Proceedings, pp.190013, 2018.

[10] F. Junliang, W. Lifeng, Z. Fucang, C. Huanjie, Z. Wenzhi, W. Xiukang, Z. Haiyang, "Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China," Renewable and Sustainable Energy Reviews, pp.186–212, 2019.

[11] N. Anikó, K. Csaba, B. Daróczy, A. Benczúr, G. Milics, J. Nagy, E. Harsányi, AJ. Kovács, M. Neményi, "Application of spatio-temporal data in site-specific maize yield prediction with machine learning methods," vol. , pp. 1–19, 2021.

[12] D. Jianhua , W. Lifeng, L. Xiaogang, F. Cheng, L. Menghui, Y. Qiliang, "Simulation of daily diffuse solar radiation based on three machine learning models," Computer Modeling in Engineering & Sciences, vol.123, pp. 49–73, 2020.

[13] T. Chen, C. Guestrin, "XGBoost : A Scalable Tree Boosting System," eprint arXiv :1603.02754v3 [cs.LG], jun. 2016.

[14] J. Friedman, "Greedy Function Approximation : A Gradient Boosting Machine", The Annals of Statistics, Vol. 29, No. 5 , pp. 1189-1232, oct. 2001.

[15] A. Saksham, A. Mohammad, R. Yasser, S. Mahdi, "Data analysis of grid-connected solar setup and regression based predictive models," pp.1–5, 2020.

[16] O. Paul, "Méthode d'apprentissage automatique appliquées au provisionnement ligne à ligne en assurance non-vie," vol.148, pp.148–162, 2017.

[17] Nogueira, F. "bayesian-optimization," https://pypi.org/project/bayesian-optimization/, 2022-01-27.

[18] Dmlc. "XGBoost Parameters," https://xgboost.readthedocs.io/en/latest/parameter.html, 2022-01-27.

[19] A. Aneseh, D. Sumit, R. Christopher, N. Olfa, P.Dan, "Robot Failure Mode Prediction with Explainable Machine Learning," IEEE International Conference on Automation Science and Engineering, pp.61–66, 2020.

[20] E. Alexis, "La valeur de Shapley-comment individualiser le résultat d'un groupe," Institut National de la Statistique et des Études Économiques, pp.53, 2012.

[21] A. Mohiuddin, S. Raihan, I. Syed, "The k-means algorithm: a comprehensive survey and performance evaluation," Electronics, pp.1295, 2020, Multidisciplinary Digital Publishing Institute.

[22] OJ. Oyelade, O. Olufunke, Obagbuwa, C. Ibidun, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," arXiv preprint arXiv:1002.2425, 2010.

[23] Bony, S., Colman, R., Kattsov, V., Allan, R., Bretherton, C., Dufresne, J., Hall, A.,Hallegatte, S., Holland, M., Ingram, W., Randall, D., Soden, B., Tselioudis, G.,and Webb, M. "How well do we understand and evaluate climate change feedback processes", Journal of Climate ,pp.3445–3482, 2006.