

Evidence, limitations and implications for renewable energy planning of daily global solar irradiation in tropical climate zones of Cameroon, predicted by machine learning

Wilson Nathan Ndackssi · Modeste Kameni Nematchoua

Received: 4 June 2026 / Accepted: 18 September 2026

© The Author(s) 2026

Abstract

Accurate solar irradiance prediction is a crucial step for the efficient planning and deployment of solar energy systems, particularly in tropical regions like Cameroon where solar resources remain largely untapped. Despite growing global interest in solar energy prediction using machine learning techniques, this remains a significant challenge. Similar studies exist in Cameroon, but they remain few compared to those in Asia, Europe, or North Africa. In contrast, our study stands out due to its complementary geographic coverage, the inclusion of solar vortex mass (SVM), and the in-depth analysis of predictive meteorological variables (particularly cloud cover). This study aims to help fill this gap by predicting daily global solar irradiance on horizontal surfaces in five Cameroonian cities Bertoua, Douala, Garoua, Maroua, and Yaoundé characterized by relatively diverse climatic conditions. Four machine learning algorithms were used for this purpose: Linear Regression, Decision Tree, Random Forest, and Support Vector Machine. Meteorological parameters including mean air temperature, relative humidity, precipitation, wind speed, wind direction, day of the year, atmospheric pressure, and cloud cover were used as input variables for training the models. The dataset was obtained from NASA's POWER database over 27 years and covers the period from 1999 to 2026. Model performance was evaluated using three statistical indicators: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). The results show that RMSE values range from 0.58 to 0.77 kWh/m², MAE values from 0.40 to 0.59 kWh/m², and R^2 values from 0.52 to 0.69, across all cities and models. Among the four algorithms evaluated, the Random Forest model consistently achieved the best predictive performance in all cities, followed by the Support Vector Machine, the Decision Tree model, and the linear model. These results suggest that the Random Forest algorithm is the most suitable model for predicting global daily solar irradiation in Cameroon and can serve as a preliminary selection tool for assessing solar energy resources in the region, while enriching the scientific literature on solar irradiation modeling in Central Africa.

Keywords Global solar irradiation · Prediction · Machine learning · Cameroon · Renewable energy

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



1 Introduction

In the current socioeconomic context, global energy demand has steadily increased with industrial and technological development. Initially met by traditional sources such as coal and biomass, this demand subsequently shifted towards oil, natural gas, hydropower, and nuclear energy as scientific advancements allowed. However, fossil fuels are being consumed at a much faster rate than their natural renewal, and although their exact depletion timeframe remains uncertain, their long-term availability is limited, making the transition to sustainable and renewable energy systems inevitable. This transition is also driven by environmental concerns, including mitigating climate change and reducing greenhouse gas emissions. Among renewable energy sources, solar energy is one of the most promising due to its abundance, sustainability, and minimal environmental impact a particularly important resource in Africa, where solar irradiance is high and relatively stable throughout the year.

In Cameroon, the energy mix remains largely dominated by hydroelectricity (62%), followed by thermal power plants (24%) and gas-fired power plants (14%), while solar energy contributes only about 0.1% to national electricity production [35]. Despite being a crude oil producing country, Cameroon remains heavily dependent on imported refined petroleum products, with a significant increase in imports recorded in recent years [35], highlighting the urgent need to diversify the national energy portfolio and make better use of local renewable resources. The country possesses considerable untapped solar potential in this regard; however, its effective exploitation requires accurate knowledge of solar radiation at ground level, which remains limited due to the scarcity of measurement stations and incomplete data collection [35, 50]. Solar radiation, emitted by the Sun in the form of electromagnetic waves, is indeed influenced by several atmospheric and geographical parameters which make its estimation complex and highly variable in space and time [25, 28], which directly affects the design, sizing and performance evaluation of photovoltaic systems. This limitation in Cameroon is part of a broader, well-documented problem across the African continent, where satellite and meteorological data play a central role in compensating for the lack of ground-based measurements. Nwokolo and Ogbulezie [45, 46], in their reviews on solar radiation in Africa, emphasize that direct measurement remains unavailable in most locations on the continent, making the use of satellite data and empirical models essential for characterizing solar resources. This importance is also highlighted by Nwokolo et al. [47] in Nigeria, and by Tesfay et al. [63] in Tigray (Ethiopia), whose comparison of satellite data and meteorological measurements confirms that combining the two sources provides a more reliable and comprehensive assessment of local solar potential, particularly in regions where ground stations remain scarce. These same satellite and meteorological data also form the basis for hybrid models combining artificial intelligence, statistical methods and physical models, which Nwokolo et al. [48] and Benatalah et al. [10] illustrate as contributing to improving the accuracy of solar radiation predictions, including when available observations are limited or incomplete. A typically African situation that only rigorous exploitation of these data sources can overcome.

Beyond mere forecasting, the exploitation of these satellite and meteorological data also opens the way to a more precise assessment of photovoltaic potential from a climate resilience perspective, as shown by Nwokolo et al. [49] and Agbor et al. [1], as well as the more applied work of Atsbeha et al. [3] and Atsbeha and Gebremariam [4] in Ethiopia, which illustrate how well-characterized solar data can be translated into concrete and resilient energy solutions. Given the importance of satellite and meteorological data for assessing solar resources and the need to utilize them through models capable of generating reliable predictions, machine learning techniques appear as a particularly suitable solution, having demonstrated a strong capacity to improve prediction accuracy when observations are limited or incomplete [29, 33]. The main objective of this study is therefore to develop and compare machine learning models for predicting daily solar irradiance in several Cameroonian cities, including Yaoundé, Douala, Garoua, Maroua, and Bertoua, using meteorological satellite data to address the knowledge gap regarding local solar resources.

2 Literature review

Over the past few decades, global solar irradiance prediction has employed a wide range of methodological approaches, from simple empirical formulations to advanced deep learning techniques, including classical statistical methods, machine learning, hybrid approaches, and, more recently, methods based on climate projections. Each of these method families relies on distinct principles, has its own advantages and limitations, and has been applied in diverse geographical and climatic contexts, notably in Africa and Cameroon, where several recent studies have enriched each of the six categories.

Solar irradiance prediction is an area where these techniques have shown particularly promising performance, notably machine learning methods and artificial intelligence techniques (Ağbulut et al. 2021; Guermoui et al. 2020; Obiora et al. 2021; Wang et al. 2016). The main objective of this work is to improve the accuracy of solar irradiance estimation to facilitate the optimal design, sizing, and operation of solar energy systems. This section reviews previous work, grouping it by methodological category, before situating the present study within these different approaches.

2.1 Empirical methods

Empirical methods rely on simple mathematical relationships, generally derived from regressions between solar irradiation and easily measurable meteorological variables (sunshine duration, temperature, humidity), most notably the Ångström-Prescott model and its numerous regional variants. In Cameroon, David and Ngwa [16] assessed the global solar radiation of five representative cities (Bamenda, Douala, Yaoundé, Ngaoundéré, and Bertoua) using the linear Ångström model and nonlinear polynomial relationships, validated by several statistical indicators (MBE, MRE, RMSE, t-stat), demonstrating the relevance of this approach for mapping national solar potential at a lower data cost. Banihashemi Dehkordi et al. [7] calibrated and validated the coefficients of this same model at six stations located in arid and semi-arid climates in Iran using optimization algorithms (Harmony Search, Shuffled Complex Evolution), obtaining mean squared errors ranging from 0.82 to 2.67 MJ/m²/day. In a comparable context, Nwokolo et al. [47, 48] developed improved Ångström-Prescott models coupled with a probabilistic Gumbel model for 37 sites across Nigeria. Da Silva et al. [15], in Botucatu (Brazil), for their part, directly compared the Ångström-Prescott empirical model to machine learning techniques (SVM, ANN): their results show that the empirical model remains more accurate than ANNs, but less accurate than SVM, which illustrates the pivotal position of this family of methods in relation to more recent approaches.

2.2 Statistical methods

Classical statistical approaches, such as simple or multiple linear regression, have also been used to predict solar radiation. Gunasekaran et al. (2021) showed in India that a basic linear regression model could be significantly improved by adding explanatory meteorological variables. Similarly, Kameni Nematchoua, Orosa, and Afaifia (2022) compared a linear model (LM) with five machine learning approaches (DT, SVM, DL, RF, GBT) for 27 European capitals, with the SVM model proving to be the most effective overall ($R^2 = 0.775$ for Paris); this study thus provides a solid statistical benchmark for evaluating the contribution of more complex methods.

2.3 Machine learning

The majority of the studies reviewed rely on machine learning, with a notable evolution over time. Early work largely employed artificial neural networks (ANNs) due to their ability to model the nonlinear relationships between meteorological variables and solar radiation. Sözen and Arcaklıoğlu [59] achieved high accuracy in Turkey, while Benghanem et al. [11] confirmed the effectiveness of ANNs in Saudi Arabia. These results were

subsequently extended to other climatic contexts by Behrang et al. [9], Rahimikhoob [52], Bilgili and Özgören [12], Leva et al. [31], and Pierro et al. (2017), consolidating the robustness of ANNs for this task.

A second wave of research has focused on kernel methods and ensemble approaches such as Random Forest (RF), Support Vector Machines (SVM/SVR), Boosting, Bagging, and Decision Trees (DT), which generally offer better generalization than classical Anomaly Networks (ANNs). Kazem et al. [29] demonstrated the effectiveness of support vector regression in the context of sparse data (Oman), while Fan et al. (2019) showed the superiority of RF and SVM models over ANNs in China. Hassan et al. [26], in North Africa (Tunisia, Algeria, Morocco), and Moungnutou Mfetoum et al. [41], in Cameroon (Bamenda), have also confirmed the value of ensemble approaches in complex climatic environments. More recently, Mutavhatsindi et al. [43], in South Africa, have also shown the potential of machine learning techniques for solar radiation forecasting. In Cameroon, Fohagui et al. [23] conducted a more recent study of five cities Bamenda, Bertoua, Ebolowa, Maroua, and Yaoundé—comparing ten machine learning and deep learning approaches: artificial neural networks (ANNs), feed-forward networks (FNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), LSTMs, linear regression (LR), k-nearest neighbors (KNNs), random forests (RFs), gradient boosting machines (GBMs), and decision trees (DTs). Their results reveal high performance, with coefficients of determination (R^2) ranging from 0.718 to 0.937 depending on the city and the model considered.

2.4 Deep learning

Deep learning architectures, particularly short- and long-term recurrent networks (LSTMs) and convolutional neural networks (CNNs), can capture complex temporal and spatial dependencies that classical machine learning models struggle to represent. In Cameroon, Molu et al. [37] developed an LSTM model for solar radiation forecasting in Douala, achieving an RMSE of 0.47 W/m² and an MAE of 5.28 W/m², surpassing reference approaches. Douswekreo et al. [19] applied an LSTM network to the northern region of Cameroon (Far North, North, and Adamawa) to simultaneously predict solar irradiance and temperature for solar electric vehicle applications, achieving an RMSE of 20.9 W/m² on 24-h forecasts from satellite data. Elsewhere, Michael et al. [34] proposed an LSTM model whose hyperparameters are optimized using a Bayesian approach, significantly improving the accuracy of short-term forecasts. Azizi et al. [5] developed a deep learning model for multi-stage and multivariate long-term solar radiation and temperature forecasting. Finally, Taha et al. [62], in Zafarana, Egypt, showed that deep architectures (CNN, LSTM, hybrid CNN-LSTM) consistently outperform classical machine learning models.

2.5 Hybrid methods

Hybrid approaches combine several families of methods signal decomposition, statistical preprocessing, and machine or deep learning models to compensate for the limitations inherent in each individual approach. Mukhtar et al. [42] developed and compared two new hybrid neural network models for hourly solar radiation forecasting, specifically designed for regions with few meteorological stations, a central problem in several sub-Saharan African countries. Huang et al. [27] proposed a WPD-CNN-LSTM-MLP multi-branch model, combining wavelet packet decomposition and deep networks, validated at three US sites (Denver, Clark, Folsom), and outperforming single-branch architectures. Ladjal et al. [30], in Ghardaïa, Algeria, developed a CNN-LSTM hybrid model for forecasting direct normal irradiance, demonstrating superior performance compared to isolated machine learning models in an arid climate. Pal Shuvo et al. [51], in Bangladesh, combined a Hodrick-Prescott filter, a discrete wavelet transform and a particle swarm optimized support vector machine (PSO-SVM), achieving a marked improvement compared to an SVM used alone.

2.6 Climate projection methods

Climate projection methods rely on regional or global climate models (RCM/GCM) and emission scenarios (RCP, SSP), generally combined with downscaling techniques, to estimate the evolution of solar radiation in the medium and long term. In West Africa, a region similar to Cameroon's climate context, Danso et al. [14] used a set of CMIP6 models under the SSP5-8.5 scenario to show that regional photovoltaic potential could decrease by nearly 12% in the southern part of the region by the end of the twenty-first century, mainly due to a decrease in surface solar radiation linked to changes in cloud cover. At the global scale, Dutta, Chanda, and Maity [20] assessed, using a multi-model set of CMIP6 models under several SSP scenarios, significant spatial variations in future photovoltaic potential. At the regional scale, Wu et al. [65] analyzed, using multi-RCM dynamic disaggregation driven by CMIP5 models, the evolution of solar potential in China under the RCP2.6, RCP4.5, and RCP8.5 scenarios. Similarly, Ferreira, Reboita, and Ribeiro [22] assessed, using a set of CMIP6 models, the future solar potential in South America under the SSP2-4.5 and SSP5-8.5 scenarios, projecting an increase of 1 to 6% in Brazil and Chile. These methods differ from the previous ones in their forward-looking time horizon, oriented towards long-term adaptation rather than short-term operational forecasting.

2.7 Synthesis and positioning of the study

Table 1 summarizes all the studies identified, classified by methodological category. This organization confirms a clear predominance of machine learning approaches in the existing literature, while also showing that the five other categories empirical, statistical, deep learning, hybrid, and climate projection. Those categories are also documented, including for the Cameroonian and West African context which now has references in each of the six categories. We have [16] for empirical, Fohagui Cyrille et al. (2026) and [41] for machine learning, [37] and [19] for deep learning, [42] for hybrid, and [14] for regional climate projection. This comprehensive overview allows us to precisely situate the present study within the state of the art and to justify the methodological choices made.

3 Method

3.1 Study sites

Cameroon is a country located in Central Africa, along the west coast of the African continent. Due to its geographical position near the equator, it experiences a tropical climate characterized by relatively high temperatures and significant spatial variability in rainfall. According to climatological studies, Cameroon is divided into several climatic zones, ranging from humid equatorial conditions in the southern regions to semi-arid and Sahelian conditions in the northern regions [38, 39]. Coastal cities are generally hot and very humid, while mountainous regions experience relatively cooler temperatures. Conversely, northern regions such as the Far North are characterized by high solar exposure, low humidity, and long dry seasons. This climatic diversity makes Cameroon a relevant case study for solar energy applications and modeling. Because of its equatorial location, the country benefits from significant solar potential. The annual sunshine duration is estimated at approximately 2500 h per year, with an average solar irradiance of around 1825 kWh/m² [18]. These values are consistent with global satellite datasets such as NASA POWER and PVGIS, which confirm the high availability of solar energy in the region [44, 61]. This strong solar potential makes Cameroon a promising area for the development of photovoltaic energy.

In this study, five cities were selected to represent the country's main climatic zones: Yaoundé, Douala, Garoua, Maroua, and Bertoua. These cities were chosen to capture Cameroon's climatic diversity and to assess the

Table 1 Summary of studies by methodological category

References	Methodological category	Location	Model	Performance
David and Ngwa [16]	Empirical	Cameroon (Bamenda, Douala, Yaoundé, Ngaoundéré, Bertoua)	Linear Ångström and polynomial relations	Statistical validation (MBE, MRE, RMSE, t-stat)
Da Silva et al. [15]	Empirical and ML comparison	Botucatu (Brazil)	AP vs SVM vs ANN	SVM > AP > ANN
Banihashemi Dehkordi et al. [7]	Empirical	Iran (6 arid/semi-arid stations)	Calibrated Ångström-Prescott (SCE/HS)	RMSE = 0.82–2.67 MJ/m ² /day
Nwokolo et al. [47, 48]	Empirical	Nigeria (37 sites)	Improved Ångström-Prescott and Gumbel	Good accuracy, generalized models
Gunasekaran et al. (2021)	Statistical	India	Linear regression	Improved by weather variables
Kameni et al. (2022)	Statistics and machine learning	27 European countries (France, Belgium, etc.)	LM / RF / DT / SVM / GBT	R ² = 0.382–0.985, ARE = 0.114–6.321, RMSE = 0.145–2.126 MJ/m ² /d
Sözen and Arcakhoğlu [59]	Machine learning	Türkiye (12 cities)	ANN (MLP)	MAPE ≤ 3.83%, R ² = 0.998
Benghanem et al. [11]	Machine learning	Al-Madinah (Saudi Arabia)	ANN (MLP)	MAPE = 2.21%, R = 0.946
Behrang et al. [9]	Machine learning	Dezful (Iran)	ANN	MAPE = 5.21%, R ² = 0.9957
Rahimikhoob [52]	Machine learning	Ahvaz (Iran)	ANN	R = 0.97, RMSE = 2.53 MJ/m ² /d
Bilgili and Özgören [12]	Machine learning	Adana (Türkiye)	ANN	MAPE = 9.23%, R ² = 0.9508
Linares-Rodríguez et al. [32]	Machine learning	Spain / Egypt	ANN / SVR / RF	RF: R ² = 0.980
Kazem et al. [29]	Machine learning	Oman	SVR	Good performance, scarce data
Hassan et al. [26]	Machine learning (set theory)	Tunisia/ Algeria/Morocco	RF / Boosting / Bagging	R ² = 0.869–0.973
Leva et al. [31]	Machine learning	Italy	ANN	MAE < 15%
Pierro et al. (2017)	Machine learning	Italy	ANN	RMSE = 7–7.5%
Liu et al. (2018)	Machine learning	China	SVM / ANN	MRE = 11.61%
Fan et al. (2019)	Machine learning	China	SVM / RF	R ² up to 0.933
Mutavhatsindi et al. [43]	Machine learning	South Africa	RF / SVR / DT / ANN	RF and SVR > linear models
Moungnutou Mfétoum et al. [41]	Machine learning (set theory)	Cameroon (Bamenda)	Set-theoretic ML models	High precision
Fohagui Cyrille et al. (2026)	Machine learning/deep learning	Cameroon (Bamenda, Bertoua, Ebolowa, Maroua, Yaoundé)	ANN/LR/KNN/CNN/RF/GBM/DT/LSTM/FNN/RNN	R ² = 0.718–0.937
Michael et al. [34]	Deep Learning	-	LSTM Bayesian optimized	Clear improvement vs. classic LSTM
Azizi et al. [5]	Deep Learning	-	Multi-stage/multivariate deep learning	Reliable long-term forecast
Taha et al. [62]	Deep Learning	Zafarana (Egypt)	CNN/LSTM/CNN-LSTM hybrid	DL > ML (short and long term)
Molu et al. [37]	Deep Learning	Douala (Cameroon)	LSTM	RMSE = 0.47 W/m ² , MAE = 5.28 W/m ²
Douswrekro et al. [19]	Deep Learning	Northern Cameroon (Far North, North, Adamawa)	LSTM (irradiation + temperature)	RMSE = 20.9 (24 h, satellite data)
Huang et al. [27]	Hybrid	Denver/Clark/Folsom (USA)	WPD-CNN-LSTM-MLP	Surpasses single-branch models
Ladjal et al. [30]	Hybrid	Ghardaia (Algeria)	CNN-LSTM hybrid	Performance > ML models only
Pal Shuvo et al. [51]	Hybrid	Dhaka/Chittagong (Bangladesh)	HP-Filter + DWT + PSO-SVM	Clear improvement vs. SVM alone
Mukhtar et al. [42]	Hybrid	Authors affiliated with Cameroon (Univ. Yaoundé I)	Two hybrid neural networks	Comparison of two hybrid architectures
Dutta et al. [20]	Climate projection	Global	CMIP6 Multi-Model Set	Marked spatial variations in PV potential

Table 1 (continued)

References	Methodological category	Location	Model	Performance
Wu et al. [65]	Climate projection	China	Multi-RCM downscaling (CMIP5, RCP2.6/4.5/8.5)	Projected increase (central/eastern China)
Ferreira et al. [22]	Climate projection	South America	CMIP6 assembly (SSP2-4.5/5-8.5)	+ 1 to 6% solar potential (Brazil/Chile)
Danso et al. [14]	Climate projection	West Africa	CMIP6 (SSP5-8.5) package	PV potential down by up to ~ 12% (southern part of the region)

robustness of machine learning models under different weather conditions. Table 2 presents the geographic coordinates and altitudes of the selected study sites, as provided by the National Institute of Cartography of Cameroon [58].

3.2 Data collection

In our study, data collection is crucial to ensuring the quality and relevance of the machine learning models we will develop. We used representative and diverse daily data, such as temperature (T2M), relative humidity (RH2M), wind speed (WS), wind direction (WD), precipitation (PRECTOTCORR), cloud cover (CLOUD_AMT), atmospheric pressure (PS), and day of the year (DOY) as explanatory variables, and global solar irradiance (ALL-SKY_SFC_SW_DWN) as the target variable. This data was obtained from the NASA POWER website [44] and covers the period from January 1, 1999, to May 30, 2026, corresponding to the geographical location of our various study sites.

3.3 Data preprocessing

Once collected, data are generally in raw form and require preprocessing before they can be used to train machine learning models. This step aims to transform the initial data into a clean, structured, and consistent set, suitable for applying different prediction algorithms. In this study, preprocessing is a prerequisite for training linear regression (LM), decision tree (DT), random forest (RF), and support vector machine (SVM) models. Appropriate preprocessing helps reduce inconsistencies in the data and adequately prepare the input variables for the models under consideration [8]. This step comprises several key phases.

The first step involves a thorough examination of the dataset to understand its structure, content, and quality. The data were downloaded in CSV format and underwent a preliminary check and cleaning process. The initial dataset contained 10,044 observations for each of the variables studied. However, the examination of the dataset revealed two distinct types of anomalies. First, 32 missing values per city were detected (coded –999 by NASA POWER), exclusively for the cloud cover variable (CLOUD_AMT) and concentrated in the last 32 days of the study period, due to the publication delay of the CERES SYN1deg satellite product. These observations were removed by truncation at a common cutoff date (May 30, 2026), ensuring a strictly identical period for all five cities.

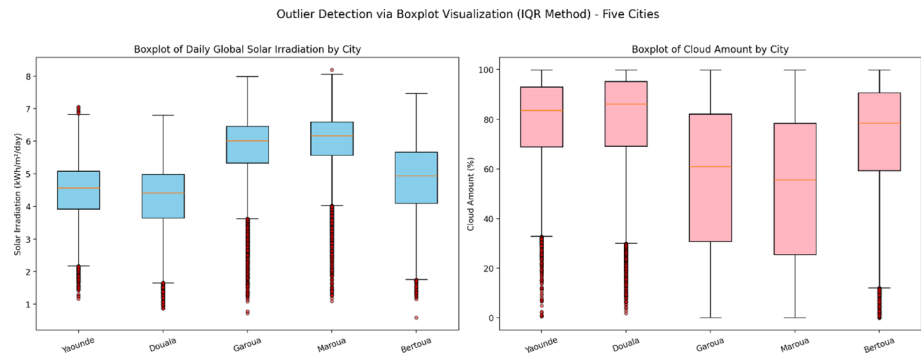
Secondly, outlier detection was performed independently of the model predictions to avoid circular bias. Figure 1 shows the box plot obtained using the IRQ method.

An initial exploratory visualization using boxplots, based on the interquartile range (IQR) method, was performed for each city. This method classifies as an outlier any observation deviating by more than 1.5 times the interquartile range of quartiles Q1 or Q3. However, when applied to data exhibiting strong seasonality, this approach proved unsuitable: in cities with a Sudano-Sahelian climate (Garoua, Maroua), it identified up to 5.7% and 6.2% of observations as outliers, respectively, corresponding in reality to rainy season days, which are statistically infrequent throughout the year but perfectly normal for this period. The classic IQR method does not distinguish between a value that is rare within the year and a value that is truly abnormal for its season. To overcome this limitation, a detection method based on a seasonal z-score was chosen. For each day of the year (DOY), the

Table 2 Spatial location of selected cities in Cameroon

Regions	Cities	Latitude	Longitude	Altitude
EAST	Bertoua	4.577° N	13.685° E	680 m
COASTAL	Douala	4.051° N	9.768° E	103 m
NORTH	Garoua	9.308° N	13.398° E	287 m
EXTREME north	Maroua	10.595° N	14.325° E	418 m
CENTER	Yaounde	3.848° N	11.502° E	589 m

Fig. 1 Outlier Detection box plot visualization for 5 cities



climatological mean $\bar{x}_{DOY}$ and the standard deviation σ_{DOY} were calculated over all available years (1999–2026), according to the sample standard deviation formula (Bessel correction, division by $n - 1$):

$$\sigma_{DOY} = \sqrt{\left(1 / (n - 1)\right) \times \sum_{i=1, n} (x_i - \bar{x}_{DOY})^2} \quad (1)$$

where σ_{DOY} represents the standard deviation of the day-of-year (DOY) values, n the number of observations, x_i the individual observed values, and $\bar{x}_{DOY}$ the mean of the DOY values.

Each observation was then standardized against its own seasonal reference:

$$z_i = (x_i - \bar{x}_{DOY}) / \sigma_{DOY} \quad (2)$$

where z_i represents the standardized value (z-score) of the i -th observation. An observation was considered an outlier when $|z_i| > 3$, a threshold corresponding to a deviation of more than three standard deviations from the normal for the same day of the year over the entire study period. This approach has the advantage of comparing each observation to its own seasonal reference rather than to the overall annual distribution, thus avoiding confusion between normal seasonal variability and actual meteorological anomalies. Even when data is of good quality, a preprocessing step remains necessary to ensure its reliability and consistency before use in predictive models. This step may include the detection and processing of outliers, the handling of missing values, and the normalization of variables. These operations contribute to improving the quality of the input data and can thus promote the robustness and reliability of the predictive performance of machine learning models [60]. This exploratory phase is essential for guiding subsequent preprocessing decisions and guaranteeing the integrity of the input data.

The meteorological variables considered in this study exhibit different scales and amplitudes of variation, which can affect the behavior of some machine learning algorithms. To limit the influence of scale differences between variables and promote a homogeneous representation of the data, a normalization step can be applied before model training. This operation is particularly relevant for methods sensitive to variable scale, such as support vector machines (SVMs) and linear regression (Díaz-Bedoya et al. 2025). Several normalization techniques can be considered, including min–max scaling, Z-score standardization, Gaussian transformation, and interquartile range (IQR)-based methods. The choice of transformation method can influence the stability and predictive performance of solar radiation forecasting models (El-Sebaï and Trabea 2023).

The meteorological variables used in this study have different value ranges, which can negatively impact model convergence and accuracy. Standardization in this work is applied only to the SVM and linear regression models to improve the predictive performance of the machine learning models. For the SVM model, the target variable (solar irradiance) was also standardized before training, and then the predictions were transformed back to their original scale ($\text{kWh}/\text{m}^2/\text{day}$) using inverse transformation before calculating the performance indicators.

After preprocessing, the dataset was subdivided into two distinct subsets to allow for the training and evaluation of machine learning models. Separating training and test data allows for the evaluation of model performance

on observations not used during training. This approach thus provides a more objective estimate of their ability to generalize to new data [64].

In the case of time series, data separation must also preserve chronological order to avoid introducing future information into the training phase. In this study, a time partition of 70% for training and 30% for testing was adopted. The first 70% of observations, corresponding to the period from January 1, 1999, to March 9, 2018, were used to train the four models considered: support vector machine (SVM), linear regression (LM), decision tree (DT), and random forest (RF). The remaining 30%, covering the period from March 10, 2018, to May 30, 2026, were retained as test data for the final evaluation of predictive performance.

This distribution provides the models with a sufficient amount of historical data to identify relationships between meteorological variables and solar irradiation, while reserving an independent portion of the dataset to assess their ability to make predictions based on observations not used during training. The choice of the proportion between training and test data depends in particular on the size of the dataset and the objectives of the study [6, 17].

3.4 General structure of the study

Machine learning enables computer systems to recognize patterns and make decisions based on accumulated experience with data. Supervised learning is one of the fundamental branches of machine learning, where a model is trained on labeled data to make predictions. In this context, supervised learning algorithms play a central role by allowing computer systems to learn from examples to predict global solar irradiance. In this study, four machine learning algorithms were used: Linear Modeling (LM), Decision Tree (DT), Random Forest (RF), and Support Vector Machines (SVM).

3.5 Machine learning algorithms

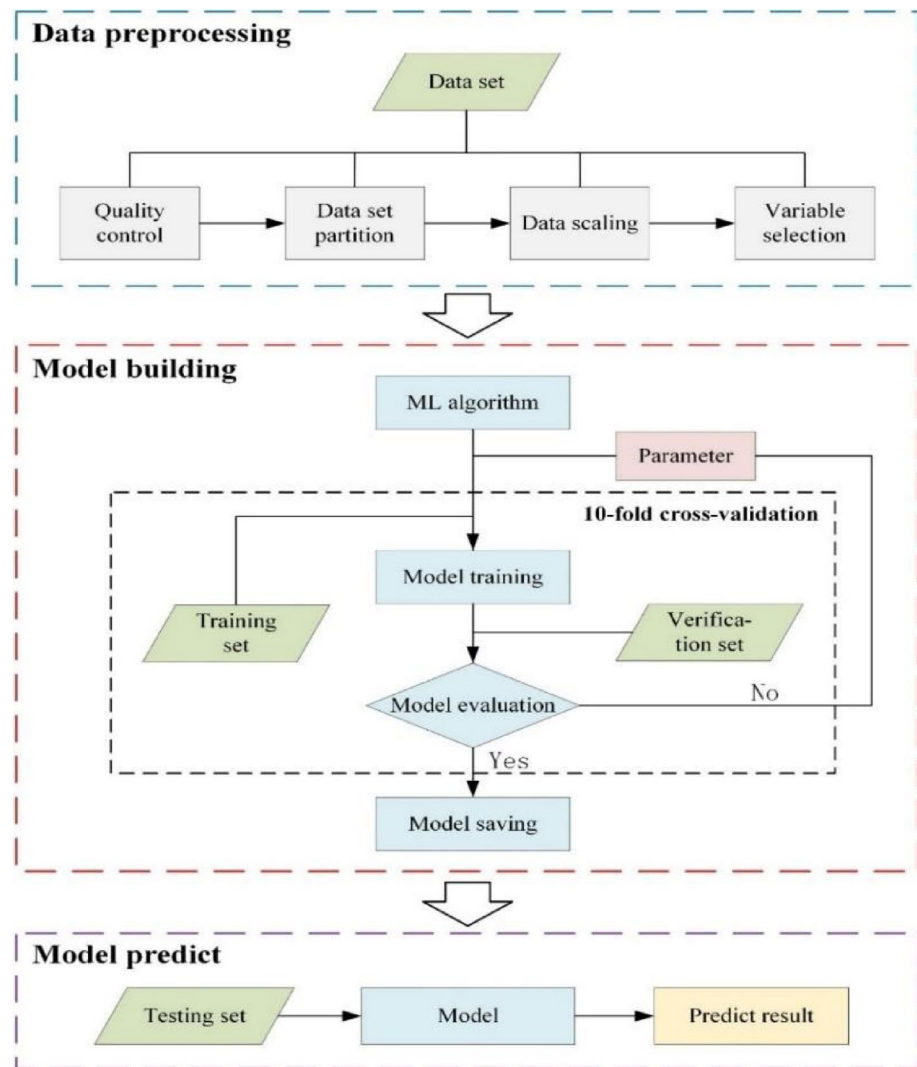
Machine learning is a branch of artificial intelligence that encompasses methods enabling algorithms to automatically identify relationships and patterns in data without requiring the explicit definition of all decision rules. After preprocessing, the data can be used to build and train machine learning models, and then to evaluate their performance on independent data. The quality of predictive performance depends, in particular, on the relevance of the selected variables, the quality of the data, and the configuration of the learning process. In this study, the machine learning approach presented in Fig. 2 is structured around three main steps: data preprocessing, model building and training, and prediction generation and evaluation. This approach is part of the general process of implementing machine learning techniques for solar radiation forecasting [27].

3.5.1 Linear regression

Linear regression is a widely used statistical and machine learning method for establishing a relationship between a dependent variable, also called the target variable, and one or more independent or explanatory variables. It involves estimating a linear function that best describes the relationship between the variables based on available observations [13, 24]. When several explanatory variables are considered simultaneously, multiple linear regression allows for estimating the combined influence of these variables on the target variable. The model is based on the assumption that a linear combination of the explanatory variables accurately represents the average trend of the dependent variable. The model coefficients, which constitute the unknown parameters, are estimated from the available data to minimize the discrepancy between observed and predicted values [24, 53].

In this study, multiple linear regression was chosen due to the presence of several meteorological variables likely to explain variations in global solar irradiance. The multiple linear regression model considered can therefore be expressed as follows:

Fig. 2 Machine learning model used for solar radiation estimation



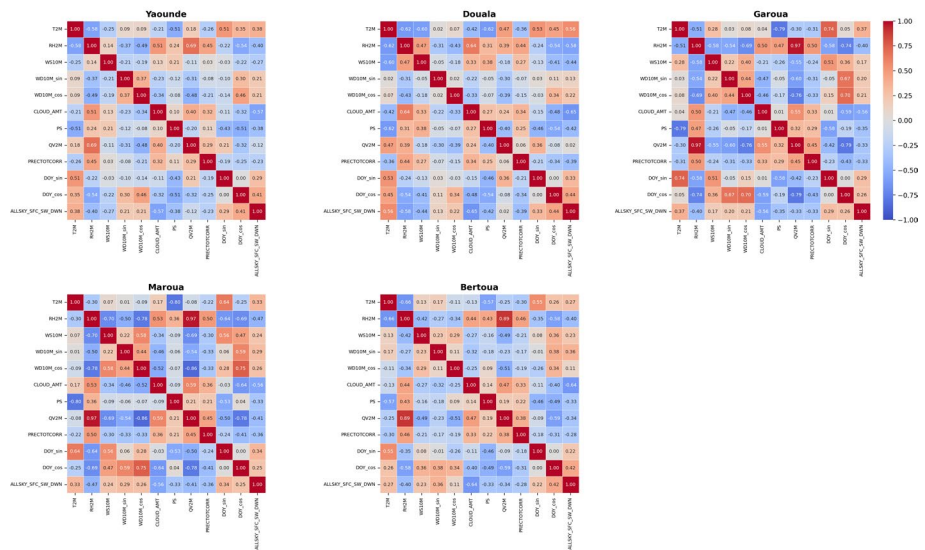
$$y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_n * X_n + \varepsilon \quad (3)$$

where y represents the dependent variable, $X_1, X_2, \dots, X_n$ the independent variables, β_0 the intercept, $\beta_1, \beta_2, \dots, \beta_n$ the regression coefficients associated with each explanatory variable, and ε the error term. The coefficients β_0 and ε are estimated by fitting the model to the training data. The output parameter corresponds to the data we wish to predict. A correlation analysis was performed to examine the relationships between meteorological variables and global solar irradiance. This analysis helps identify the variables with the strongest associations with the target variable and, consequently, to better guide the selection of explanatory variables used in the prediction models (Islam et al. 2025). Pearson correlation coefficients between each input variable and global solar irradiance were calculated for the five cities studied. Figure 3 presents the Pearson correlation matrices between the explanatory and target variables for the five study cities (Yaoundé, Douala, Garoua, Maroua, and Bertoua).

3.5.2 Decision Tree

Decision Trees (DTs) are a machine learning method widely used in artificial intelligence and data analysis. They rely on a tree structure to make decisions or predictions based on a set of successive conditions. For quantitative data, building a decision tree involves selecting the most appropriate explanatory variables and separation

Pearson Correlation Matrices - Five Cities (Outliers Removed)



City	Max_depth	Min_samples_split	Min_samples_leaf	Max_features
Douala	7	2	20	
Yaoundé	7	2	20	0.8
Bertoua	7	2	20	0.8
Garoua	5	2	10	0.8
Maroua	5	2	10	

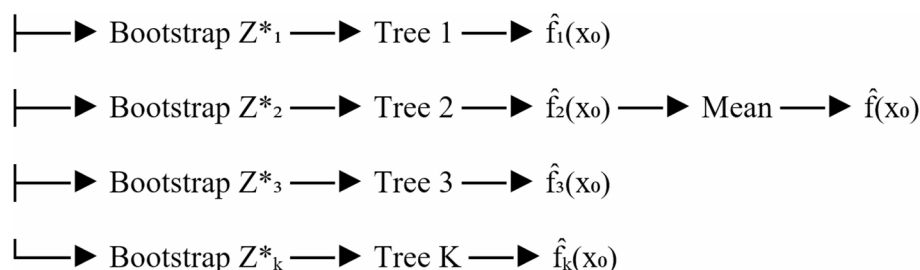
To assess the contribution of localized modeling, we compared the performance of a single decision tree trained on all five cities (global model) to that of five distinct decision trees, each trained and optimized independently on data specific to a single city (individual models). For each approach, a hyperparameter search using cross-validation (GridSearchCV, 5-fold) was performed to select the optimal configuration (maximum depth, minimum number of samples per leaf and per split, number of variables considered at each split). Table 3 below presents the hyperparameters found.

3.5.3 Random Forest

This aggregation strategy reduces the variance associated with individual trees, thus limiting their susceptibility to overfitting. By combining predictions from several diverse trees, the random forest can offer a favorable trade-off between bias and variance, while also improving the stability of the predictions. This property is particularly interesting for solar radiation forecasting, where meteorological variations and fluctuations in input

variables can introduce significant variability into the data. Applications of the random forest to solar radiation prediction have notably demonstrated its potential to simultaneously exploit several meteorological variables, such as air temperature, relative humidity, and wind speed (Nematchoua et al. 2022; Konfuzio 2023). The overall algorithmic flow of the Random Forest model can be summarized as follows:

Training set:



Thanks to its overall architecture, its ability to handle high-dimensional data, and its robustness to outliers and missing data, Random Forest has established itself in the machine learning literature as one of the most widely used and reliable algorithms for predicting solar irradiance (Konfuzio 2023). Furthermore, the random selection of variables during the construction of each tree allows for the decoupling of individual predictions from different trees, thus enhancing the overall performance of the model (Ho 1995).

In this work, a search for hyperparameters was conducted, followed by cross-validation. The following Table 4 presents a summary of the optimal hyperparameters obtained for each city studied.

We observe that in the cities of Douala, Yaoundé, and Bertoua, the trees composing the forest are deeper (max_depth=7) and the leaves more constrained (min_samples_leaf=20), while in the cities of Garoua and Maroua, the trees are shorter (max_depth=5) and the leaves less constrained (min_samples_leaf=10). This shows that even when constructing several hundred trees on different bootstrap subsamples, cross-validation systematically converges on deeper trees for the cities of Douala, Yaoundé, and Bertoua. This reinforces the hypothesis of intrinsic data complexity.

3.5.4 Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning method widely used to solve classification and regression problems (Burgess 1998; Cortes and Vapnik 1995). It is based on the principle of minimizing structural risk, which aims to simultaneously control the learning error and the complexity of the model in order to improve its ability to generalize to unobserved data (Vapnik 1995).

In the context of regression problems, the method is known as Support Vector Regression (SVR). This approach allows the construction of a prediction function by finding a solution that keeps errors within a tolerance range while controlling the model's complexity. Through the use of kernel functions, SVR can also represent nonlinear relationships between explanatory variables and a continuous target variable, making it particularly well-suited to forecasting problems with complex relationships [57].

The rbf kernel is systematically selected, both globally and individually, confirming the presence of non-linear relationships between meteorological variables (particularly cloud cover) and solar irradiance. The value of $C=10$ returns for four of the five cities, reflecting a similar bias/variance trade-off from one city to another;

Table 4 Hyperparameters for each city for random forests

City	Max_depth	Min_samples_split	Min_samples_leaf	Max_features
Douala	7	2	20	Sqrt
Yaoundé	7	2	20	Sqrt
Bertoua	7	2	20	0.8
Garoua	5	2	10	0.8
Maroua	5	2	10	Sqrt

only Bertoua stands out with a lower C (1), suggesting a higher tolerance for training errors, potentially linked to a more marked local climate variability (Table 5).

Support Vector Regression (SVR) relies in particular on the use of an ε -insensitive loss function, which defines a tolerance zone around the observed values. Errors whose magnitude remains less than or equal to this margin ε are not penalized by the loss function, while deviations exceeding this zone contribute to model optimization. This formulation allows SVR to limit the influence of small variations or noise on the learning process and to focus model fitting on the most significant deviations [57]. Given an input vector x , SVR projects the data into a higher-dimensional variable space using a nonlinear transformation $\phi(x)$, and constructs a linear regression function in this space:

$$f(x) = w \cdot \phi(x) + b \quad (4)$$

where $f(x)$ represents the output function, w the weight vector, x the input vector, $\phi(x)$ a transformation function applied to x , $\cdot$ the dot product between w and $\phi(x)$, and b the bias term. This projection allows the SVR to efficiently model the nonlinear relationships between the input meteorological variables and the target output. The SVR model was configured using the radial basis function (RBF) kernel, defined as follows:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (5)$$

where $K(x_i, x_j)$ denotes the similarity between observations (x_i) and (x_j), $\|x_i - x_j\|^2$ is the squared Euclidean distance between them, (γ) is the kernel parameter controlling the width of the Gaussian function, and $(\exp)$ denotes the exponential function. The radial basis function (RBF) kernel is one of the kernel functions commonly used with support vector regression (SVR) to represent nonlinear relationships in high-dimensional spaces [54]. Its use is particularly relevant for solar irradiance forecasting, the variability of which results from complex and potentially nonlinear relationships between different meteorological variables. In this context, factors such as temperature, relative humidity, sunshine duration, and cloud cover can interact in complex ways and influence observed irradiation levels. The association of solar radiation potential (SVR) with an RBF kernel has been investigated in several studies on solar radiation forecasting. Mohammadi et al. [36] notably used SVR to predict global solar radiation over a horizontal surface, while Shamshirband et al. [56] also demonstrated the value of this approach for solar radiation estimation. These results highlight the potential of SVR for modeling nonlinear relationships between meteorological variables and solar radiation and explain its use in various solar energy forecasting applications.

3.6 Evaluation indicators

It is important to choose appropriate evaluation measures depending on the context and objectives of the machine learning problem. In this study, several measures are applied, such as MAE (mean absolute error), RMSE (root mean square error), and R^2 (coefficient of determination), to compare the performance of these predictive models. These different statistical measures, their corresponding equations, and brief descriptions are presented in Table 6.

Table 5 Hyperparameters for each city for the SVM algorithm

City	Kernel	C	Gamma	Epsilon
Douala	Rbf	10	0.1	0.5
Yaounde	Rbf	10	0.1	0.5
Bertoua	Rbf	1	Auto	0.5
Garoua	Rbf	10	0.1	0.5
Maroua	rbf	10	0.1	0.5

Table 6 Statistical measures

Indicators	Equations	Description and references
Mean absolute error (MAE)	$\frac{1}{n} \sum_{i=1}^n y_i - x_i $	MAE is a long-term prediction error indicator, where values close to 0 indicate better model performance [49]
Root mean square error (RMSE)	$\sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}}$	RMSE reflects short-term prediction and is always positive; values close to 0 indicate good accuracy [21]
Coefficient of determination R^2	$1 - \frac{\sum_{j=1}^n (y_i - x_i)^2}{\sum_{i=1}^n (x_i - \bar{x}_i)^2}$	It takes values between [0, 1], and good model performance is indicated by an R^2 value close to 1 (Rehman 1998)

The coefficient of determination R^2 is a statistical measure used to assess the proportion of the variance of the dependent variable that can be explained by the independent variables included in the model. It thus serves as an indicator of the model's goodness of fit to the observed data. A high R^2 value indicates that the model explains a significant portion of the variability of the target variable [40].

It provides a measure of the overall quality and predictive performance of a statistical model, with values ranging from 0 to 1, interpreted as follows:

- An R^2 value < 0.3 indicates negligible or very low explanatory power.
- A value of $0.3 \leq R^2 < 0.5$ indicates a small effect size.
- A value of $0.5 \leq R^2 < 0.7$ indicates a moderate effect size.
- An R^2 value ≥ 0.7 indicates a large effect size, reflecting high predictive accuracy.

4 Results and discussion

4.1 Correlation analysis between meteorological variables and solar irradiation

This section examines the statistical relationships between the meteorological input variables and daily global solar irradiation across the five studied cities. Figure 3 presents the Pearson correlation matrices obtained for each city, including inter-variable correlations relevant to the assessment of multicollinearity.

Cloud cover (CLOUD_AMT) consistently exhibits the highest absolute correlation across all five cities. The second most influential predictor, however, varies by climatic zone: seasonality (DOY_cos) ranks second in Yaoundé and Bertoua, both located in the equatorial southern region, whereas relative humidity (RH2M) takes precedence in Douala, Garoua, and Maroua. This contrast suggests that solar irradiation variability in the humid equatorial cities is more strongly tied to the annual seasonal cycle, while in the coastal and Sudano-Sahelian cities, moisture content plays a more direct role irrespective of season. A strong correlation was also observed between relative humidity (RH2M) and specific humidity (QV2M), initially considered as a candidate input variable, with coefficients reaching 0.97 in Maroua and Garoua. Given this near-collinearity and its potential to destabilize the estimated coefficients of the linear regression model, QV2M was excluded from the final set of input variables used for model training.

4.2 Performance of machine learning algorithms

This work evaluates the prediction of daily solar irradiance in five Cameroonian cities using four machine learning algorithms, and we discuss their performance using three indicators commonly found in the literature. Table 4 presents the different numerical values of the metrics for the various cities studied. For each model developed, the correlation coefficient R^2 varies between 0.45 and 0.69 (Table 7).

The table confirms a very stable performance ranking across the five cities:

Table 7 Comparative analysis of the performance of these algorithms across the five cities

Cities	Metrics	Linear model	Decision tree	Random forest	Support vector machines
Douala	MAE	0.50	0.46	0.43	0.43
	RMSE	0.66	0.62	0.58	0.6
	R ²	0.53	0.58	0.64	0.61
Bertoua	MAE	0.59	0.57	0.52	0.51
	RMSE	0.76	0.73	0.68	0.69
	R ²	0.49	0.52	0.58	0.57
Maroua	MAE	0.53	0.45	0.40	0.39
	RMSE	0.73	0.64	0.60	0.60
	R ²	0.52	0.63	0.69	0.67
Garoua	MAE	0.53	0.49	0.46	0.44
	RMSE	0.77	0.70	0.66	0.67
	R ²	0.49	0.58	0.63	0.62
Yaounde	MAE	0.50	0.46	0.42	0.43
	RMSE	0.65	0.61	0.58	0.58
	R ²	0.45	0.51	0.56	0.55

Linear model < Decision tree < SVM < Random forest

The linear model consistently performs the worst (R^2 between 0.45 and 0.53), which is expected: linear regression assumes an additive and linear relationship between meteorological variables and solar irradiance, whereas the physics of the phenomenon (cloud cover threshold effect, interactions between humidity and temperature, cyclical seasonality) is inherently non-linear. This model serves as a baseline here useful: it allows us to measure the real contribution of non-linear methods, which gain between +0.06 and +0.17 points of R^2 depending on the city.

Forest and SVM these two non-linear methods are very similar. They exhibit almost equivalent performance across all cities (R^2 difference often less than 0.02), and vie for first place depending on the city and metric considered: Random Forest is the best among the four models tested on Maroua and Garoua, while the SVM's MAE is sometimes marginally inferior (Douala, Bertoua, Maroua). This similarity is explained by the fact that the two methods, although very different in their construction, are both capable of capturing complex non-linear relationships.

The decision tree alone still positions itself between the linear model and the two most advanced methods: it already captures a good part of the non-linearities (gain of +0.05 to +0.11 of R^2 compared to the linear model), but remains penalized by its intrinsic instability (only one tree is sensitive to sampling noise), which the aggregation of the Random Forest precisely corrects.

The hierarchy between cities is confirmed; regardless of the model chosen, a constant ranking of difficulty is found, whatever the model:

Bertoua emerged as the most difficult city to model (R^2 between 0.49 and 0.58 depending on the model), slightly ahead of Yaoundé. This suggests that Bertoua and Yaoundé share a similar high level of difficulty.

We observe a consistency between the RMSE/MAE ratio. Indeed, they remain stable for the same model across cities at approximately 1.3 to 1.4, which indicates a fairly homogeneous error distribution, without the presence of isolated extreme errors that would cause the RMSE to explode relative to the MAE, thus showing that the treatment of outliers carried out upstream with the Z-score was effective.

Non-linear ensemble or kernel methods (Random Forest, SVM) clearly outperform linear or single-tree approaches, with a comparable gain in robustness between them; the difficulty of modeling solar irradiance varies significantly according to the climatic zone, with the cities of the center and east (Bertoua, Yaoundé) remaining structurally more difficult to predict than those of the northern Sahel (Maroua, Garoua), regardless of the algorithm used.

4.3 Summary of R^2 by city and by model

Of the five cities studied, the Random Forest model consistently obtained the highest R^2 (0.56 in Yaoundé to 0.69 in Maroua). This perfect regularity, without any exceptions, is a strong argument for retaining this model as the final reference model for the study: it constitutes a structural advantage that holds true regardless of the local climatic context.

The table perfectly illustrates the expected logic in machine learning: the more a model is able to capture complex and non-linear relationships, the better its R^2 : Linear Regression (0.45–0.53) → Decision Tree (0.51–0.63) → SVM (0.55–0.67) → Random Forest (0.56–0.69). This progression is almost monotonic for each city taken individually, which reinforces the interpretation that the performance gain comes from the model's ability to represent the non-linearity of the physical phenomenon (irradiance being strongly dependent, and non-linearly, on cloud cover), and not from a statistical artifact.

4.4 Importance of variables in the random forest algorithm

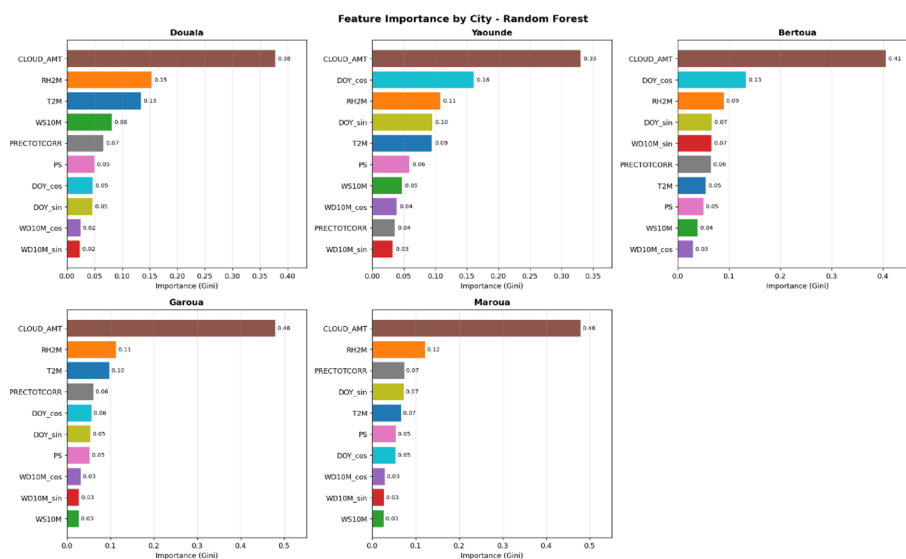
The Fig. 4 below helps to understand the influence of the input variables on the model.

Cloud cover emerges as the most significant variable in all five cities, with an importance ranging from 0.33 (Yaoundé) to 0.48 (Garoua, Maroua), representing between one-third and one-half of the total predictive capacity. Of The model relies on this single variable. This result is physically logical: solar irradiance at ground level depends directly on the fraction of radiation blocked by clouds before reaching the surface. No other variable comes close, confirming that cloud cover is the primary physical factor in the phenomenon under study.

Relative humidity consistently ranks second or third (0.09 to 0.15), with a contribution three to four times smaller than CLOUD_AMT. It likely acts as an indirect indicator of cloud cover not fully captured by CLOUD_AMT. alone (humid air favoring cloud formation), as well as by a proper diffusion/absorption effect of radiation by atmospheric water vapor.

Seasonality (DOY_sin/DOY_cos), for its part, has a variable weight depending on the city. Its contribution differs significantly from one city to another. In Yaoundé and Bertoua, DOY_cos Reaching a high significance (0.16 and 0.13 respectively), its position within the year is a strong and consistent signal. In Douala, Garoua, and Maroua, seasonality has a much smaller impact (0.05–0.07). Daily cloud cover appears to follow a more irregular pattern, less closely aligned with the annual cycle. The cloud cover-season relationship is better captured by

Fig. 4 Feature importance by city for random forest



the model in Yaoundé/Bertoua, but this does not result in a better overall R^2 . Seasonality is significant there, but insufficient on its own for accurate prediction, hence the lower R^2 despite this signal.

Temperature regularly ranks in third or fourth position (0.06–0.13), with a more pronounced contribution in Douala (0.13). It likely captures part of the effect of radiation itself as well as correlated seasonal climate variations.

Precipitation (0.04–0.07), surface pressure (0.05–0.06), and wind speed (0.03–0.08) each make a modest but real contribution. Precipitation likely acts as a complementary signal to CLOUD_AMT. for intense rainfall episodes (partly justifying the error peaks observed during these events), while pressure and wind capture finer synoptic variations.

The cyclical components of wind direction consistently show the lowest values (0.02 to 0.07). Wind direction has an indirect and tenuous physical link with solar irradiance their low weight confirms that they mainly provide marginal information, consistent with their secondary role already observed in the simple decision tree.

The hierarchy of variables is remarkably stable from one city to another. Cloud cover is the most important factor, humidity is the second most important, but seasonality is the primary factor. Regional differentiation is more pronounced in areas with a biphasic equatorial climate (Yaoundé, Bertoua) than in areas with a more homogeneous regime (Douala, Garoua, Maroua). This hierarchy confirms, at the model level itself, the physical relevance of the variables selected during preprocessing.

4.5 Prediction results

Across the five cities, the error curve oscillates in a balanced way around the zero line, with a fairly regular alternation between black zones, proving that the model overestimates. The actual irradiance and red zones show that the model underestimates it. This alternation, rather than long and unidirectional blocks, confirms the absence of structural bias already observed on the scatter plots: the Random Forest does not tend to systematically favor one direction of error, it reacts on a case-by-case basis to daily variations (Figs. 5, 6, 7, 8, 9).

The background noise, consisting of small oscillations between -0.5 and $+0.5$ kWh/m², is interrupted by isolated peaks that can reach ± 1.5 to ± 2 kWh/m². These peaks almost always correspond to abrupt breaks in the “Actual” series of the top panel, typically associated with isolated weather events (isolated thunderstorms, localized cloud cover) that the model cannot anticipate since it only has daily averages and no information on intra-day dynamics.

Fig. 5 Actual solar irradiation data, prediction result and error magnitude for Maroua using the Random forest algorithm

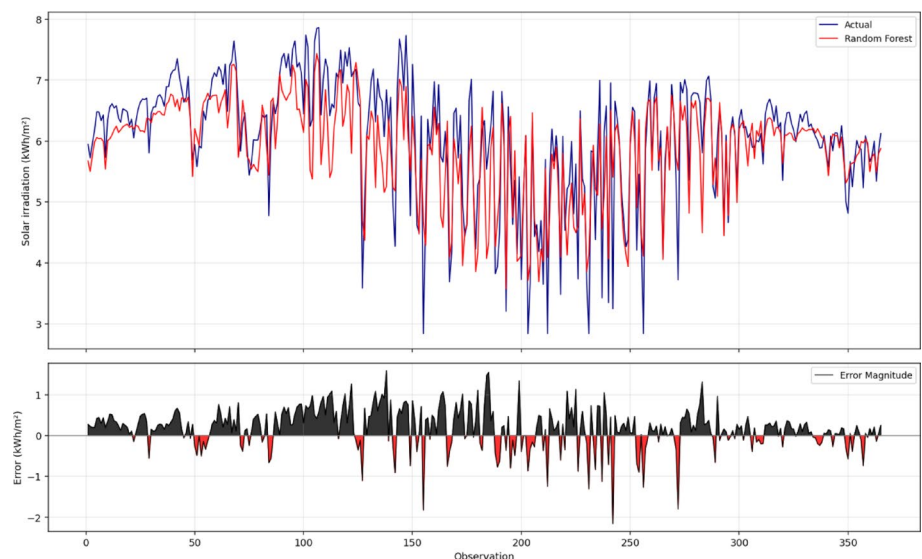


Fig. 6 Actual solar irradiation data, prediction result and error magnitude for Bertoua using the Random forest algorithm

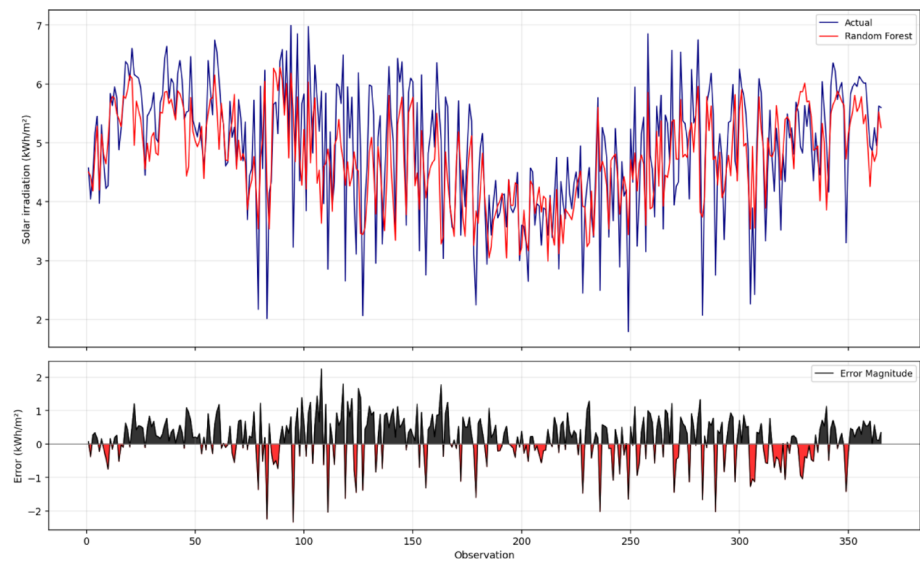
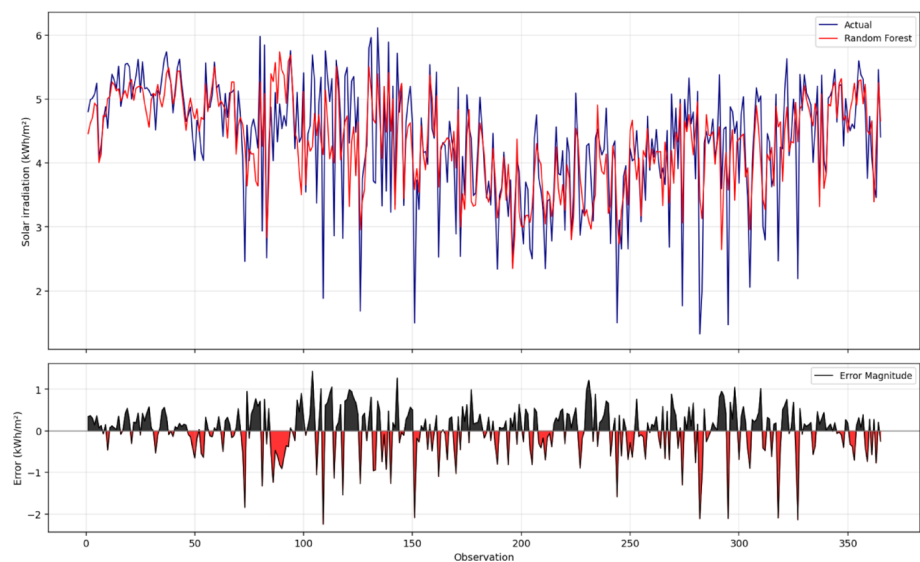


Fig. 7 Actual solar irradiation data, prediction result and error magnitude for Douala using the Random forest algorithm



For several cities, notably Maroua and Garoua, the highest error peaks are clearly concentrated in the second half of the year (beyond the 150–200 observation period), a period that generally corresponds to the rainy season in these regions. This is consistent with the previously put forward hypothesis: cloud variability associated with rainfall is more difficult to model than a stable, clear sky during the dry season, where irradiance follows a more predictable trajectory.

Visually, the amplitude of the error peaks follows the same hierarchy as the RMSE values in the table: cities with the lowest R^2 values, such as Bertoua and Yaoundé, exhibit more frequent and higher error peaks than Maroua or Douala, where deviations exceeding $\pm 1 \text{ kWh/m}^2$ remain less frequent. This parallel between the error diagram and the aggregate metrics is a good sign of the analysis's internal consistency: both the R^2 /RMSE reading and the detailed reading via the error time series are coherent.

Random Forest produces generally small and unbiased errors, but remains vulnerable to sudden and punctual breaks in cloud cover, particularly during the rainy season, which remain the main source of residual error for the five cities studied (Fig. 10).

Fig. 8 Actual solar irradiation data, prediction result and error magnitude for Garoua using the Random forest algorithm

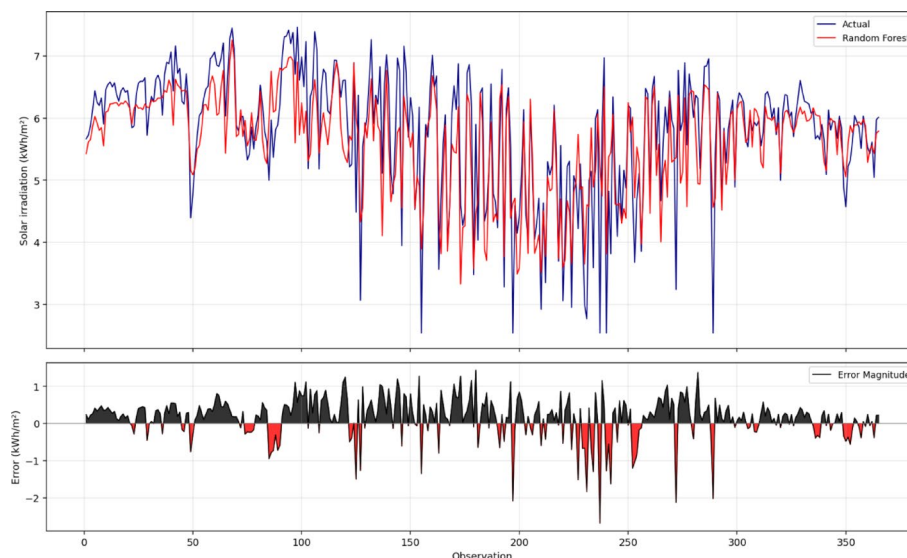
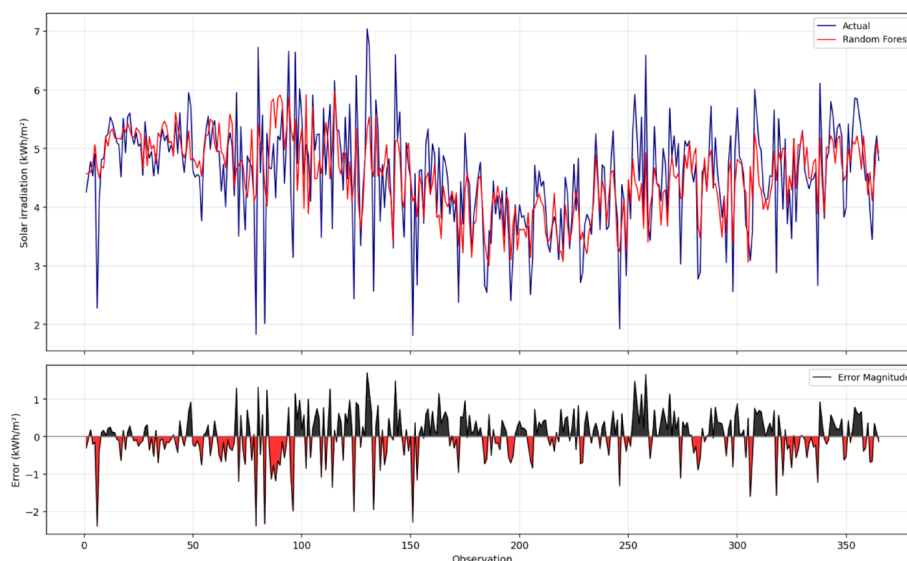


Fig. 9 Actual solar irradiation data, prediction result and error magnitude for Yaoundé using the Random forest algorithm

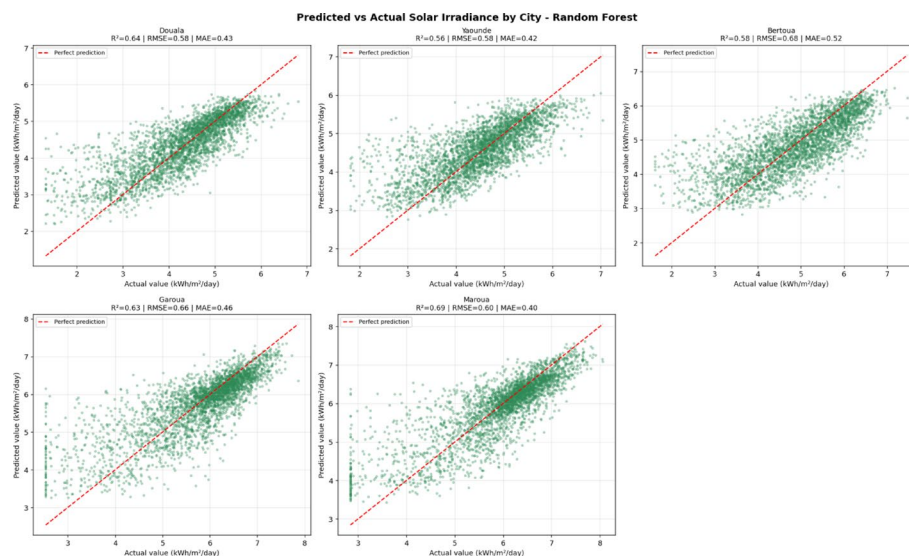


Across the five graphs, the scatter plot generally follows the red regression line, visually confirming the R^2 values in the table (0.56–0.69 depending on the city). No city shows systematic drift; the model does not exhibit structural bias, and it underestimates and overestimates in a relatively balanced way.

vertical column of points is observed in the lower left corner of each graph, corresponding to very low actual values, around 2–3 kWh/m²/day, associated with significantly higher predictions, 3–5 kWh/m²/day. These are typically exceptionally cloudy or rainy days, where actual irradiance drops sharply—an event that the Random Forest model, built on averages from many trees, tends to smoothing rather than capturing precisely. This is a known limitation of aggregation methods: they reduce variance at the cost of a reduced ability to predict rare extreme values.

Bertoua point cloud is clearly the most spread out around the diagonal, consistent with its lowest R^2 in the group (0.58) and its highest RMSE (0.68). This increased dispersion reflects a greater intrinsic variability in irradiance at Bertoua, which current meteorological variables struggle to fully explain. It is probably linked to more unpredictable convective cloud cover in this forested area.

Fig. 10 Scatter graph of each city according to machine learning algorithms: Douala, Yaoundé, Bertoua, Garoua, and Maroua



Conversely, Maroua ($R^2 = 0.69$) and Douala ($R^2 = 0.64$) show the most concentrated point clouds around the diagonal, particularly in the central part of the distribution (average irradiance values, the most frequent). The model is therefore most reliable there for current climatic conditions.

It is interesting to note that irradiance levels differ from one city to another: Garoua and Maroua reach predicted/actual values of up to 8 kWh/m²/day, compared to approximately 7 kWh/m²/day for Douala, Yaoundé, and Bertoua. This confirms, regardless of the model's performance, a higher solar potential in the northern areas (drier climate, less cloud cover on average).

Overall, the Random Forest model demonstrates satisfactory predictive performance across the five cities, with Maroua showing the best agreement between measured and predicted values. These results confirm that the Random Forest algorithm is the model in our study that best captures the non-collinearity relationship for predicting daily global solar irradiance.

4.6 Comparison with existing studies

The study most directly comparable to ours is that of Fohagui et al. [23], which also predicts daily solar irradiance in five Cameroonian cities (Bamenda, Bertoua, Ebolowa, Maroua, Yaoundé) using NASA POWER data, notably testing a Random Forest model common to both studies. Three cities are common to both samples (Bertoua, Maroua, Yaoundé), while our study additionally covers Douala (humid coastal climate) and Garoua (Sudanian-Sahelian climate), two climatic zones absent from their study, thus strengthening the geographical and climatic coverage of Cameroon.

Across the three cities, their Random Forest model achieved an R^2 of 0.907 in Yaoundé, 0.890 in Bertoua, and 0.892 in Maroua, compared to 0.56, 0.58, and 0.69 respectively in our study for the same algorithm. This difference, obtained with an identical model, cannot be explained by a different methodological choice at the algorithm level; rather, it reflects differences in data preprocessing and the set of variables used.

First, their preprocessing replaces missing values with the mean and negative or outliers with zero, smoothing out some of the extreme variability of the target signal. Our study, conversely, preserved this actual variability by relying on anomaly detection based on a seasonal z-score (threshold $|z| > 3$), a more conservative approach that maintains the intrinsic difficulty of the prediction task at the cost of a more modest R^2 . Second, it is noteworthy that Fohagui et al. [23] do not include cloud cover (CLOUD_AMT) among their input variables, whereas our Pearson correlation analysis consistently identifies this variable as the most strongly correlated predictor of solar irradiance in the five cities studied, suggesting that the performance gap does not stem from a lack of input

information on our part. Finally, their longer study period (41 years, 1980–2021, compared to 27 years, 1999–2026, in our case) provides them with a larger training volume, which is particularly favorable to the stability of the model estimates.

This comparison with identical algorithms suggests that the conservative choice to retain the real extreme values rather than smoothing them, combined with a shorter observation period, explains most of the R^2 difference observed in common cities, rather than a weakness in the modeling itself. Moreover, the inclusion of Douala and Garoua allows our study to document two climatic zones of Cameroon not yet covered by the existing literature.

5 Conclusion

This study presented a comparative analysis of four machine learning algorithms—Linear Regression, Decision Tree, Random Forest, and Support Vector Machine—for predicting daily global solar irradiance in five Cameroonian cities: Bertoua, Douala, Garoua, Maroua, and Yaoundé. The performance of each algorithm was evaluated using three indicators: the root mean square error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2).

The results obtained from the comparative analysis consistently demonstrated that the Random Forest algorithm outperformed the other models in all cities and for all evaluation indicators, recording the lowest RMSE and MAE values as well as the highest R^2 scores. In particular, the city of Maroua produced the best prediction results with an R^2 of 0.69 and an RMSE of 0.58, while Yaoundé proved to be the most difficult city to model, recording the lowest R^2 value (0.52) and the highest RMSE (0.60) for all algorithms. The linear model consistently recorded the weakest performance in all cities, making it the least suitable algorithm for predicting solar irradiance in this study. The decision tree model and Support Vector Machines have shown moderate and sometimes competitive performance, especially in cities such as Garoua and Maroua.

The analysis of the point clouds further confirmed these results, showing that the predicted values generated by the Random Forest model do not have a structural bias; they underestimate and overestimate in a relatively balanced way.

This study confirms that the Random Forest algorithm is the most suitable model for predicting global daily solar irradiation in Cameroon amongst the four algorithm, offering a robust preliminary selection tool that can support decision-making in the renewable energy sector, particularly for the planning and optimization of solar energy systems.

Regarding future work, three main directions have been identified. First, the use of more robust datasets combining satellite and meteorological data could significantly improve the predictive performance of models, reduce prediction errors, and allow for a more accurate characterization of solar irradiation patterns across the various climatic zones of Cameroon. Second, exploring deep learning models such as long short-term memory (LSTM) networks, convolutional neural networks (CNNs), and hybrid architectures would be of great interest to observe their behavior and predictive capacity to better capture the non-collinearity between variables for this type of tropical and equatorial climate, characterized by high variability and complex atmospheric conditions. Thirdly, propose policy recommendations with advice for careful deployment on solar PV sites, mini-grids, rural electrification, and solar pumping.

Acknowledgements The authors would like to thank NASA for providing the meteorological and solar radiation data used in this study. Statements relating to ethics, consent to participation and consent to publication: not applicable.

Author contributions Author Contributions: Wilson Nathan NDACKSSI: Conceptualization, Methodology, Software, Data curation, Formal analysis, Writing—original draft. Modeste KAMENI NEMATCHOUA: Supervision, Validation, Writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding This research received no external funding.

Data availability The data used in this study were obtained from NASA's POWER database and are publicly available at <https://power.larc.nasa.gov/>

Declarations

Ethics approval and consent to participate Ethical approval was not required for this study because it did not involve human participants, human subjects, or the collection of personal or identifiable information. The study was conducted exclusively using publicly available meteorological and solar radiation data obtained from the NASA POWER database. This study did not involve human participants or the collection of personal data. The analysis was based exclusively on publicly available meteorological and solar radiation data from the NASA POWER database.

Consent for publication Not applicable. This study does not contain any individual-level personal data, identifiable information, or individual images requiring consent for publication.

Competing interests The authors declare that they have no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Agbor ME, Udo SO, Ewona IO, Nwokolo SC, Ogbulezie JC, Amadi SO. Potential impacts of climate change on global solar radiation and PV output using the CMIP6 model in West Africa. *Clean Eng Technol*. 2023;13:100630. <https://doi.org/10.1016/j.clet.2023.100630>.
2. Asana. Decision tree analysis. 2024. <https://asana.com/en/resources/decision-tree-analysis>
3. Atsbeha AK, Berhe TA, Gebru MT. Modeling and analysis of flat plate solar air collector with phase change materials for extended surface, drying application. *Discov Energy*. 2026;6:4. <https://doi.org/10.1007/s43937-025-00113-4>.
4. Atsbeha AK, Gebremariam GG. Techno-economic feasibility analysis of an off-grid solar-wind-battery hybrid renewable energy system for the Kara Adishuho community, Northern Ethiopia. *Renew Sustain Energy Technol*. 2026;2(1):6. <https://doi.org/10.53941/rset.2026.100003>.
5. Azizi N, Yaghoubirad M, Farajollahi M, Ahmadi A. Deep learning based long-term global solar irradiance and temperature forecasting using time series with multi-step multivariate output. *Renew Energy*. 2023;206:135–47. <https://doi.org/10.1016/j.renene.2023.01.102>.
6. Bakirci K. Correlations for estimation of daily global solar radiation with hours of bright sunshine in Turkey. *Energy*. 2009;34(4):485–501.
7. Banihashemi Dehkordi SN, Bakhtiari B, Qaderi K, Ahmadi MM. Calibration and validation of the Angstrom–Prescott model in solar radiation estimation using optimization algorithms. *Sci Rep*. 2022;12:4855. <https://doi.org/10.1038/s41598-022-08744-6>.
8. Bassam A, May Tzuc O, Escalante Soberanis M, Ricalde LJ, Cruz B. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Front Earth Sci*. 2021;9:596860. <https://doi.org/10.3389/feart.2021.596860>.
9. Behrang MA, Assareh E, Ghanbarzadeh A, Noghrehabadi AR. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. *Sol Energy*. 2010;84(8):1468–80.
10. Benatallah M, Bailek N, Bouchouicha K, Sharifi A, et al. Solar radiation prediction in Adrar, Algeria: a case study of hybrid extreme machine-based techniques. 2024. <https://doi.org/10.4028/p-VH0u4y>
11. Benghanem M, Mellit A, Alamri SN. ANN-based modeling and estimation of daily global solar radiation data: a case study. *Energy Convers Manag*. 2009;50(7):1644–55. <https://doi.org/10.1016/j.enconman.2009.03.035>.
12. Bilgili M, Özgören M. Daily total global solar radiation modeling from several meteorological data. *Meteorol Atmos Phys*. 2011;112(3–4):125–38. <https://doi.org/10.1007/s00703-011-0137-9>.

13. Brownlee J. Linear regression for machine learning. Machine Learning Mastery. 2019. <https://machinelearningmastery.com/linear-regression-for-machine-learning/>
14. Danso DK, Anquetin S, Diedhiou A, Lavaysse C, Hingray B, Raynaud D, et al. A CMIP6 assessment of the potential climate change impacts on solar photovoltaic energy and its atmospheric drivers in West Africa. *Environ Res Lett*. 2022;17(4):044016. <https://doi.org/10.1088/1748-9326/ac5a67>.
15. Da Silva MBP, Escobedo JF, Rossi TJ, Santos CM, Silva SHMG. Performance of the Angstrom-Prescott model (AP) and SVM and ANN techniques to estimate daily global solar irradiation in Botucatu/SP/Brazil. *J Atmos Solar Terr Phys*. 2017;160:11–23. <https://doi.org/10.1016/j.jastp.2017.04.001>.
16. David A, Ngwa NR. Global solar radiation of some regions of Cameroon using the linear Angstrom and non-linear polynomial relations (Part I): model development. *Int J Renew Energy Res*. 2013;3(4):984–92.
17. Dietz C, et al. Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets. ResearchGate. 2024.
18. Djaowe S. Decentralized rural electrification by individual photovoltaic systems: case of the locality of Ntui-Esson. 2014.
19. Douswekreo S, Ndoumbe J, Baba A, Fedotova M, Offole F, Essola D. Solar irradiation and temperature prediction using LSTM neural network and solar energy potential in Northern Cameroon for solar electric vehicle application. *J Eng*. 2024. <https://doi.org/10.1155/je/1536889>.
20. Dutta R, Chanda K, Maity R. Future of solar energy potential in a changing climate across the world: a CMIP6 multi-model ensemble analysis. *Renew Energy*. 2022;188:819–29. <https://doi.org/10.1016/j.renene.2022.02.023>.
21. Fan J, Wang X, Wu L, Zhang F, Bai H, Lu X, et al. New combined models for estimating daily global solar radiation based on sunshine duration in humid regions: a case study in South China. *Energy Convers Manag*. 2018;156:618–25. <https://doi.org/10.1016/j.enconman.2017.11.085>.
22. Ferreira GWS, Reboita MS, Ribeiro JGM. Assessment of the solar energy potential over South America estimated by CMIP6 models in the present and future climate. *J Environ Earth Sci*. 2024;6(2):110–43.
23. Fohagui FCV, Koholé YW, Wankouo Ngoueu CA, Noutchogouin Tedom D, Tchuen G. Prediction of Cameroon's global solar radiation using deep learning and machine learning algorithms. *J Atmos Sol Terr Phys*. 2026. <https://doi.org/10.1016/j.jastp.2026.106733>.
24. Freedman DA. Statistical models: theory and practice. Cambridge University Press; 2009.
25. Gairaa K, Khellaf A, Messlem Y, Chellali F. Estimation of the daily global solar radiation based on Box-Jenkins and ANN models: a combined approach. *Renew Sustain Energy Rev*. 2016;57:238–49. <https://doi.org/10.1016/j.rser.2015.12.111>.
26. Hassan MA, Khalil A, Kaseb S, Kassem MA. Exploring the potential of tree-based ensemble methods in solar radiation modeling. *Appl Energy*. 2017;203:897–916. <https://doi.org/10.1016/j.apenergy.2017.06.104>.
27. Huang X, Li Q, Tai Y, Chen Z, Zhang J, Shi J, et al. Hybrid deep neural model for hourly solar irradiance forecasting. *Renew Energy*. 2021;171:1041–60. <https://doi.org/10.1016/j.renene.2021.02.161>.
28. Kaldellis JK, Zafirakis D. Optimum energy storage techniques for the improvement of renewable energy sources-based electricity generation economic efficiency. *Energy*. 2011;36(4):2267–78. <https://doi.org/10.1016/j.energy.2010.06.010>.
29. Kazem H, Khatib T, Sopian K. Support vector regression for prediction of solar radiation in Oman. *Energy Convers Manag*. 2013;65:168–75. <https://doi.org/10.1016/j.enconman.2012.08.015>.
30. Ladjal B, Nadour M, Bechouat M, Hadroug N, Sedraoui M, Rabehi A, et al. Hybrid deep learning CNN-LSTM model for forecasting direct normal irradiance: a study on solar potential in Ghardaia, Algeria. *Sci Rep*. 2025. <https://doi.org/10.1038/s41598-025-94239-z>.
31. Leva S, Dolara A, Grimaccia F, Mussetta M, Ogliari E. Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. *Math Comput Simul*. 2017;131:88–100. <https://doi.org/10.1016/j.matcom.2015.05.010>.
32. Linares-Rodríguez A, Ruiz-Arias JA, Pozo-Vázquez D, Tovar-Pescador J. Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks. *Energy*. 2011;36(8):5356–65.
33. Mellit A, Kalogirou SA. Artificial intelligence techniques for photovoltaic applications: a review. *Prog Energy Combust Sci*. 2008;34(5):574–632. <https://doi.org/10.1016/j.pecs.2008.01.001>.
34. Michael NE, Hasan S, Al-Durra A, Mishra M. Short-term solar irradiance forecasting based on a novel Bayesian optimized deep long short-term memory neural network. *Appl Energy*. 2022;324:119727. <https://doi.org/10.1016/j.apenergy.2022.119727>.
35. Ministry of Water and Energy, Cameroon. National energy statistics report. 2023.
36. Mohammadi K, Shamshirband S, Anisi MH, Alam KA, Petković D. Support vector regression based prediction of global solar radiation on a horizontal surface. *Energy Convers Manag*. 2015;91:433–41. <https://doi.org/10.1016/j.enconman.2014.12.015>.
37. Molu RJJ, Dzonde Naoussi S, Wira P, Fendzi Mbasso W. Solar irradiance forecasting based on deep learning for sustainable electrical energy in Cameroon. *Int J Smart Grid*. 2023. <https://doi.org/10.20508/ijsmartgrid.v7i2.279.g261>.
38. Molua EL, Lambi CM. Climate, water and agriculture in Cameroon (IFPRI Discussion Paper). International Food Policy Research Institute. 2006a.

39. Molua EL, Lambi CM. Climatic trends in Cameroon: implications for agricultural management. *Clim Res.* 2006;30(3):255–62.
40. Moore DS, Notz WI, Fligner MA. The basic practice of statistics. 6th ed. W. H. Freeman and Company; 2013.
41. Moungnutou Mfetoum I, Koumi Ngoh S, Molu RJJ, Nde Kenfack BF, Onguene R, Dzonde Naoussi SR, et al. A multilayer perceptron neural network approach for optimizing solar irradiance forecasting in Central Africa with meteorological insights. *Sci Rep.* 2024. <https://doi.org/10.1038/s41598-024-55883-z>.
42. Mukhtar M, Oluwasanmi A, Yimen N, Qinxu Z, Ukwuoma CC, Ezurike B, et al. Development and comparison of two novel hybrid neural network models for hourly solar radiation prediction. *Appl Sci.* 2022;12(3):1435. <https://doi.org/10.3390/app12031435>.
43. Mutavhatsindi T, Sigauke C, Mbuva R. Forecasting hourly global horizontal solar irradiance in South Africa using machine learning models. *IEEE Access.* 2020;8:198872–85. <https://doi.org/10.1109/ACCESS.2020.3034690>.
44. NASA. NASA POWER: prediction of worldwide energy resources. 2024. <https://power.larc.nasa.gov>
45. Nwokolo SC, Ogbulezie JC. A comprehensive review of empirical models for estimating global solar radiation in Africa. *Renew Sustain Energy Rev.* 2017. <https://doi.org/10.1016/j.rser.2017.04.101>.
46. Nwokolo SC, Ogbulezie JC. A qualitative review of empirical models for estimating diffuse solar radiation from experimental data in Africa. *Renew Sustain Energy Rev.* 2018;92:353–93. <https://doi.org/10.1016/j.rser.2018.04.118>.
47. Nwokolo SC, Amadi SO, Obiwulu AU, Ogbulezie JC, Eyibio EE. Prediction of global solar radiation potential for sustainable and cleaner energy generation using improved Angstrom-Prescott and Gumbel probabilistic models. *Clean Eng Technol.* 2022;6:100416. <https://doi.org/10.1016/j.clet.2022.100416>.
48. Nwokolo SC, Obiwulu AU, Ogbulezie JC, Amadi SO. Hybridization of statistical machine learning and numerical models for improving beam, diffuse and global solar radiation prediction. *Clean Eng Technol.* 2022;9:100529. <https://doi.org/10.1016/j.clet.2022.100529>.
49. Nwokolo SC, Ogbulezie JC, Ushie OJ. A multi-model ensemble-based CMIP6 assessment of future solar radiation and PV potential under various climate warming scenarios. *Optik (Stuttg).* 2023;285:170956. <https://doi.org/10.1016/j.ijleo.2023.170956>.
50. Obafemi TS, Oyewole OO, Afolabi AS. Assessment of solar energy potential in Africa using satellite data and ground observations. *Renew Sustain Energy Rev.* 2019;102:105–19. <https://doi.org/10.1016/j.rser.2018.11.022>.
51. Pal Shuvo S, Pal Shibaze S, Paul G, Paul Mita M, Das C, Malakar K. A novel integration of Hodrick–Prescott filter (Hp-filter) and wavelet transform (WT) with optimized support vector machine (PSO-SVM) in predicting solar radiation. *Sci Rep.* 2025. <https://doi.org/10.1038/s41598-024-82950-2>.
52. Rahimikhoob A. Estimating global solar radiation using artificial neural network and air temperature data in a semi-arid environment. *Renew Energy.* 2010;35(9):2131–5.
53. Rencher AC, Christensen WF. Methods of multivariate analysis. 3rd ed. Wiley; 2012.
54. Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press; 2002.
55. Seal HL. The historical development of the Gauss linear model. *Biometrika.* 1967;54(1–2):1–24. <https://doi.org/10.1093/biomet/54.1-2.1>.
56. Shamshirband S, et al. Prediction of the solar radiation on the Earth using support vector regression technique. *Infrared Phys Technol.* 2015;68:179–85. <https://doi.org/10.1016/j.infrared.2014.12.006>.
57. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput.* 2004;14(3):199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>.
58. Soda. Altitude of a point. 2024. <https://www.soda-pro.com>
59. Sözen A, Arcaklioğlu E. Solar potential in Turkey. *Appl Energy.* 2005;80(1):35–45. <https://doi.org/10.1016/j.apenergy.2004.03.004>.
60. Subramanian E, Karthik MM, Krishna GP, Kumar VS, Prasath DV. Solar power prediction using machine learning. 2023. arXiv. <https://arxiv.org/abs/2303.07875>
61. Šuri M, Huld TA, Dunlop ED, Ossenbrink HA. Potential of solar electricity generation in the European Union member states and candidate countries. *Sol Energy.* 2007;81(10):1295–305. <https://doi.org/10.1016/j.solener.2006.12.007>.
62. Taha A, Makeen P, Nazih N. Short-term and long-term solar irradiance forecasting with advanced machine learning techniques in Zafarana, Egypt. *Sci Rep.* 2025. <https://doi.org/10.1038/s41598-025-24853-4>.
63. Tesfay AH, Atsbeha AK, Hailu MH. Solar resource mapping of the Tigray region, Ethiopia, based on satellite and meteorological data. *Energies.* 2025;18(19):5264. <https://doi.org/10.3390/en18195264>.
64. University of Illinois. Split test/train. Data Science Discovery. 2023. <https://discovery.cs.illinois.edu/learn/Towards-Machine-Learning/TestTrain-Split/>
65. Wu J, et al. Future projection of solar energy over China based on multi-regional climate model simulations. *Earth Space Sci.* 2022. <https://doi.org/10.1029/2021EA002207>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Wilson Nathan Ndackssi¹ · Modeste Kameni Nematchoua^{1,2}

✉ Modeste Kameni Nematchoua
Kameni.modeste@yahoo.fr; mkameni@uliege.be

¹ Laboratory of Energy, Electrical Systems and Electronics, Department of Physics, University of Yaoundé I, Yaoundé, Cameroon

² Department of Civil Engineering, University of Antsirana, Antsirana, Madagascar