Prediction of site-specific solar diffuse horizontal irradiance from two input variables in Colombia

Elieser Miranda, Jorge Felipe Gaviria Fierro, Gabriel Narváez, Luis Felipe Giraldo, Michael Bressan

Department of Electrical and Electronic Engineering, University of the Andes, Bogotá, Colombia

ARTICLE INFO

Keywords:
Diffuse horizontal irradiance
Sun position angles
Clearness index
Machine learning
Random forest
Gaussian processes

ABSTRACT

Accurate measurements of diffuse irradiance are essential to design a solar photovoltaic system. However, in-situ radiation measurements in Colombia, South America, can be limited by the costs of the implementation and maintenance of meteorological stations equipped with a pyranometer mounted on a sun tracker with a shading device, which is required to measure diffuse irradiance. Furthermore, the databases found in Colombia contain missing data, which raises the need for implementing models that are trained with very few features. In this paper, we introduce a methodology based on simple angle calculations and a regression model to predict half-hourly diffuse horizontal solar irradiance from only the measure of global horizontal irradiance and a geographic coordinate as inputs. Using measurements taken from the national solar radiation database for 6 different sites in Colombia and state-of-the-art machine learning models for regression, we validated the accuracy prediction of the proposed methodology. The results showed a prediction error ranging from 5.86 to 9.36 [W/m²], and a coefficient of determination ranging from 0.9974 to 0.9983. The data-set used along with the feature engineering process and the deep neural network model created can be found in a Github repository referenced in the paper.

1. Introduction

The solar radiation received by a photovoltaic (PV) field can be decomposed in three different parts: the direct radiation, which is the solar beam received directly from the sun in clear sky scenarios; diffuse radiation, which is the solar radiation received from the atmosphere due to scattering cloudy sky; and reflected solar radiation, which is the solar radiation received by reflection from ground, windows and puddle. The sum of these three radiations corresponds to the global radiation. Predicting global solar irradiance and its three components at a given location is crucial for the optimal design, installation, simulation, and evaluation of solar PV systems [1]. In particular, to measure diffuse horizontal irradiance ($D_h$), a pyranometer with sun blocking disk mounted on sun tracker can be a solution [2, 3]. This represents high installation and maintenance costs that in developing countries such as Colombia, South America, can be difficult to afford, especially when underprivileged communities are the ones that will benefit from PV systems. To solve this problem, many researchers have used indirect methods such as modeling techniques or satellite estimation methods that allow for predicting $D_h$ from the global horizontal irradiance ($G_h$) [4]. Meteorological variables and the clearness index ($M_t$) have been considered the most important inputs to develop solar prediction models of diffuse irradiance on horizontal surfaces [5]. In addition to these variables, sun position angles such as solar declination angle and altitude angle have been considered to improve model accuracy [6]. In recent years, machine learning algorithms have also been used to address this problematic. In [7], the researchers compared four different machine learning algorithms, including artificial neural networks, kernel and nearest-neighbor and support vector machines, to predict global solar radiation. Furthermore, in [8], the researchers used artificial neural networks and random forest to predict hourly three components of solar radiation (global horizontal irradiation, beam normal and diffuse horizontal radiation) for different time horizons. In [9], the authors used ensemble learning techniques in solar irradiance predictions. These methods included random forest, support vector regression and artificial neural networks.

In the previously mentioned papers, the authors do not account the fact that many datasets usually contain missing data, as the ones found in Colombia. For example, the Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM) database contains 194...
The main contribution of our paper is to implement a methodology to predict half-hourly diffuse irradiance. Following, we present a short review of previous work that have addressed the problem of diffuse horizontal irradiance prediction using only two variables. Several models have been proposed to estimate hourly, daily and monthly averages but not accurate for hourly predictions. Years later, between 1982 and 1990, Erbs et al. [13], Skartveit y Olseth [14], Maxwell [15], Pérez et al. [16] developed different models to determine the $f_D$ per hour, taking the Liu and Jordan model as reference. There are many empirical correlation models between $M_t$ and $f_D$ that are commonly used because of their simplicity and acceptable performance in various geographic and climatic conditions [17]. Abreu et al. [18] presented a review of 121 hourly models in which $M_t$ is the sole predictor to compute $D_h$.

### 2. Related work

Several models have been proposed to estimate hourly, daily and monthly averages of diffuse irradiance. Following, we present a short review in which we group them into three categories: empirical models, multivariate empirical models, and machine learning models.

#### 2.1. Empirical models

This models typically are polynomial functions of the clearness index ($M_t$) and the diffuse fraction ($f_D$). For example, Liu and Jordan’s work [12] reported the relationship between $M_t$ and $f_D$ using measurements from 98 stations in Canada and the United States. The mentioned model proved to be efficient for predictions with monthly averages but not accurate for hourly predictions. Years later, between 1982 and 1990, Erbs et al. [13], Skartveit y Olseth [14], Maxwell [15], Pérez et al. [16] developed different models to determine the $f_D$ per hour, taking the Liu and Jordan model as reference. There are many empirical correlation models between $M_t$ and $f_D$ that are commonly used because of their simplicity and acceptable performance in various geographic and climatic conditions [17]. Abreu et al. [18] presented a review of 121 hourly models in which $M_t$ is the sole predictor to compute $D_h$.

#### 2.2. Multivariate empirical models

This models typically are linear combinations of individual variables or pairwise multiplications of variables. For example, Reindl et al. [19] developed a multivariate empirical model using $M_t$, altitude angle, ambient temperature, and relative humidity in five European and North American locations. They obtained a reduction of 14% in the prediction error, compared with the empirical models at the same place. Afterward, Skartveit [20] improved their empirical model in 1998 by adding the hourly variance index and regional surface albedo to the set of input variables, showing that the new approach outperforms Erbs, Maxwell, Perez models. Therefore, numerous multivariate empirical models were developed in the last decades that proposed a correlation among the $f_D$ and a set of variables as $M_t$, temperature, relative humidity, declination, altitude angle, azimuth angle, albedo, and hour angle, obtaining models with better results than the empirical of a single predictor [21].

### 2.3. Machine learning models

Machine learning models are commonly used in the field of renewable energy due to their capability of learning from data without explicit programming. In this work, we have considered different machine learning algorithms to predict diffuse irradiance. The selected algorithms were trained with hourly diffuse irradiance measurements from the same locations used for the empirical models. The results of the machine learning models were compared with the empirical models and the best performing model was selected for the validation process.

#### 2.3.1. Regression models

Regression models are commonly used in the field of renewable energy due to their capability of learning from data without explicit programming. In this work, we have considered different machine learning algorithms to predict diffuse irradiance. The selected algorithms were trained with hourly diffuse irradiance measurements from the same locations used for the empirical models. The results of the machine learning models were compared with the empirical models and the best performing model was selected for the validation process.

#### 2.3.2. Neural network models

Neural network models are commonly used in the field of renewable energy due to their capability of learning from data without explicit programming. In this work, we have considered different machine learning algorithms to predict diffuse irradiance. The selected algorithms were trained with hourly diffuse irradiance measurements from the same locations used for the empirical models. The results of the machine learning models were compared with the empirical models and the best performing model was selected for the validation process.

---

**Table 1.** Nomenclature used throughout the paper.

| Symbol | Description                     |
|--------|---------------------------------|
| $G$    | global solar radiation          |
| $S$    | direct solar radiation          |
| $D$    | diffuse solar radiation         |
| $R$    | reflected solar radiation       |
| $Sh$   | direct horizontal radiation     |
| $Sh$   | diffuse horizontal radiation    |
| $Gh$   | global horizontal radiation     |
| $P$    | photovoltaic                    |
| $i_0$  | extraterrestrial radiation      |
| $fi$   | incidence angle                 |
| $fD$   | diffuse fraction                |
| $f_D$  | diffuse fraction                |
| $fD$   | diffuse fraction                |
| $GSC$  | solar constant                  |
| $MBE$  | mean bias error                 |

**Table 2.** Relevant information from the selected locations in Colombia for model validation.

| City       | Alt. [m a.s.l.] | Avg. temp. [°C] | Lat. | Long. |
|------------|-----------------|-----------------|------|-------|
| Caruru     | 185             | 28              | 1.01 | -71.3 |
| Barrancominas | 100             | 31              | 3.49 | -69.82|
| Chajal     | 65              | 32              | 1.61 | -78.54|
| Sipi       | 85              | 25.4            | 4.65 | -76.66|
| Puerto Merizalde | 11             | 26              | 3.32 | -77.42|
| Bogota     | 2630            | 14.5            | 4.60 | -74.06|
2.3. Machine learning models

To capture the relationship between \(D_h\) and predictor variables such as \(Mt\) and meteorological variables, machine learning models have been proposed as an alternative to define a wider variety of non-linear regression functions [22]. Soares et al. [23] used an artificial neural network (ANN) to estimate hourly values of \(D_h\) at the surface in São Paulo City, Brazil, using \(G_h\) and other meteorological parameters as input variables. They concluded that the atmospheric long-wave radiation used as an input improves the neural-network performance and showed that ANNs can be more accurate than by empirical models to predict \(D_h\). In a similar way, the work [24] and [25] conducted similar experiments in Egypt and India, respectively, to predict \(D_h\), reaching the same conclusion with respect to ANNs. Kaushika et al. [26] used as inputs latitude, longitude, altitude, time, month of the year, relative humidity, total rainfall, and sunshine duration to build an ANN in New Delhi City, India. The model was capable to compute \(D_h\), \(S_h\), \(G_h\) with excellent performance for both dry and wet months. Hassan et al. [27] explored the use of gradient boosting, bagging, and random forest models for regression to estimate global and diffuse irradiance. The clearness index, sunshine duration, and the maximum possible number of daylight (hour duration) were utilized as inputs. They concluded that the models presented very reliable and accurate results, despite being relatively simple.

3. Materials and methods

3.1. Dataset

Colombia is a country located at the north part of South America. The Colombian territory has its limits to the north with the Caribbean Sea, to the northwest with Panama, to the south with Ecuador and Peru, to the east with Venezuela, to the southeast with Brazil, and to the west with the Pacific. Our methodology to obtain accurate predictions of \(D_h\) is validated using measurements collected from the National Solar Radiation Database NSRDB [4]. This database includes measurements of \(G_h\) and \(D_h\) variables, taken every 30 min from January 2007 to December 2017, from six positions located at the center, west, south, and northwest of Colombia. We selected these locations because they represent the different weather and geographical conditions in Colombia.

Relevant information for each location is provided in Table 2.

Inaccurate global solar radiation values were identified based on clearness index \((Mt)\) such that the values that did not fall into the range of 0.015 < \(Mt\) < 1 were rejected [28]. Likewise, measurements were filtered according with the sunrise and sunset at values not less than 10 [W/m²].

3.2. Proposed methodology

Figure 2 shows a block diagram of the proposed methodology to predict diffuse horizontal irradiance \(D_h\) in Colombia using regression models. This process requires as first step collecting historical data of \(G_h\) at different locations, with the corresponding time stamps and their geographic coordinates. Given this information, a feature engineering is conducted in which the following variables are extracted for each time stamp: angles associated with the position of the sun relative to a plane with a given orientation such as declination \(\delta\), hour angle \(\omega\), solar azimuth \(\varphi\), and altitude angle \(h\). Using these angles, two additional angles are computed: angle of incidence \(\theta\) and zenithal angle \(\theta_z\). Those angles can be calculated easily with equations that are included in standard books on solar energy such as [29].

The angle of declination \((\delta)\) is the angle between the Earth–Sun vector and the equatorial plane and can be computed as shown in Eq. (1). Here, \(n_d\) represents the day number.

\[
\delta = 23.45 \sin \left( \frac{360 \times 284 + n_d}{365} \right) \tag{1}
\]

Legal Time (TL) can be given as

\[
TL = TU + GMT \tag{2}
\]

According to [29] and [30], the angular displacement \((\omega)\) of the Sun from the local point is defined by

\[
\omega = (TSV - 12)15. \tag{3}
\]

Table 3. Hyperparameters of the Random Forest (RF) model after training.

| City            | Num. Trees | Min. Leaf Size | Num. of Predictors |
|-----------------|------------|----------------|-------------------|
| Caruru          | 29         | 2              | 6                 |
| Barrancominas   | 10         | 5              | 7                 |
| Chajal          | 185        | 2              | 8                 |
| Sipí            | 28         | 1              | 8                 |
| Puerto Merizalde| 10         | 2              | 7                 |
| Bogotá          | 41         | 1              | 8                 |
point on the Earth, sun position can be determined by two main angles, namely, altitude angle ($h$) and azimuth angle ($a$) [30], which were respectively expressed by Eqs. (4) and (5). The variable $la$ represents the latitude of the place of interest.

$$\sin(h) = \cos(\delta)\cos(\omega)\cos(la) + \sin(\delta)\sin(la), \quad (4)$$

$$\sin(a) = \frac{\cos(\delta)\sin(\omega)}{\cos(h)}; \quad (5)$$

Duffy and Beckman [29] mentioned that the angle between the vertical and the line to the sun, that is, the angle of incidence of beam radiation on a horizontal surface is called the zenith angle. For horizontal surfaces, the angle of incidence is the zenith angle of the sun, $\theta_z$. Its value must be between 0 and 90 degrees when the sun is above the horizon. The following equation relates $\theta_z$ to angles $a$, $\delta$, and $\omega$.

$$\cos(\theta_z) = \cos(la)\cos(\delta)\cos(\omega) + \sin(la)\sin(\delta). \quad (6)$$

The solar radiation that is received on a horizontal surface located at the upper limit of the atmosphere is called extraterrestrial radiation $I_o$. The value is defined from the value of the solar constant ($G_{sc}$), that is the average of the incident energy in a surface unit ($m^2$) perpendicular to the direction of propagation of the radiation at mean earth-sun distance outside the atmosphere.

The value of $G_{sc} = 1367 \text{ [W/m}^2\text{]}$ has been adopted by the World Radiation Center (WRC) with an uncertainty of the order of 1%. Since extraterrestrial radiation varies at each time of the year in the range of ±3.3%, a simple equation with accuracy adequate for most engineering calculations is given by Spencer [31].

$$I_o = G_{sc}\left(1 + 0.33 \cos\left(\frac{360\omega}{365}\right)\right). \quad (7)$$

Also, the clearness index or cloud transmittance factor $Mt$ is computed. This index is defined as the ratio of solar radiation on a given surface compared to the extraterrestrial radiation $I_o$:

$$Mt = \frac{G_h}{I_o\cos(\theta_z)} \quad (8)$$

This parameter incorporates both light scattering and light absorption and varies between 1 and 0. When $Mt$ is close to 1, the sky is very clear and if it is close to 0, the sky is very cloudy. As a result, the clearness index may be considered as an attenuation of the atmosphere. We hypothesize that these variables contain enough information to capture the behavior of $Dh$ in the Colombian territory.

Given these variables, a regression model is then trained to capture the relationship between the extracted features and $Dh$. We used four different types of models to validate our methodology: empirical and multi-variable empirical models, which are the classic models for $Dh$ prediction; models for Gaussian process regression, tree-based ensemble methods for regression, and artificial neural networks, which are regression models for machine learning that are known to provide accurate prediction results. Following, we provide a short introduction of each method.

**Empirical (EMP) and Multi-variable Empirical (MVEMP).** The usual approach for $Dh$ prediction using empirical models is the utilization of a correlation factor named diffuse fraction defined as $fd = Dh/Gh$. Abreu [18] verified that many authors present the same model but for different location to calculate the diffuse fraction based on the clearness index ($Mt$). Those models were developed using several functional forms such as second-degree polynomial as a function of $Mt$, higher polynomial degrees, and double exponential forms. To predict $Dh$ using empirical model form (EMP), the diffuse fraction ($Dh/Gh$) was correlated with the

| City          | Num. of Trees | Min. Leaf Size |
|---------------|---------------|---------------|
| Caruru        | 15            | 1             |
| Barrancominas | 57            | 1             |
| Chajal        | 10            | 4             |
| Sipi          | 10            | 8             |
| Puerto Merizalde | 50       | 1             |
| Bogota        | 37            | 1             |

Table 4. Hyperparameters of the Ensemble Bagged Tress (EBT) after training.

Figure 3. Scatter plots of measured and predicted values of $Dh$ using EMP, MVEMP, EGPR, EBT,RF, and ANN models.
clearness index ($Mt$). On the other hand, multi-variable empirical model (MVEMP) correlates the diffuse fraction with a set of variables.

**Gaussian Process Regression (EGPR).** Gaussian process regression (GPR) models are non-parametric kernel-based probabilistic models whose properties are entirely determined by the mean and covariance functions of a real process [32]. There are a variety of kernel functions that can be selected to define the covariance function of the Gaussian processes. In this work, we chose the exponential or Gaussian kernel [33], since it has been shown that it allows for improved performance for regression compared to processes using other covariance functions.

**Ensemble methods: Random Forest (RF) and Bagged Regression Trees (EBT).** These methods have been shown to achieve a balance between bias and variance for regression [27]. RF consists of a large number of decision trees that work jointly, where each tree makes a class prediction, and the class with the most votes becomes the prediction of the model [34]. On the other hand, EBT is used for reducing the variance of the model by creating several subsets of data used to train several decision trees. The average of the predictions from the different trees is used [35].

**Artificial Neural Network (ANN).** The expression neural network has improved to involve a large category of models and learning methods, which is studied and analyzed extensively in [36]. Artificial Neural Networks (ANN) are functions whose structure is defined by the serial and parallel interconnection of basic operations [37].

### Table 5. Results for the predicting of solar $Dh$. Independent test.

| City          | Model | MBE [$W/m^2$] | RMSE [$W/m^2$] | $R^2$  |
|---------------|-------|---------------|----------------|--------|
| Caruru        | EMP   | 9.7536        | 106.2809       | 0.5092 |
|               | MVEMP | 0.9312        | 37.0466        | 0.9373 |
|               | EGPR  | 0.8621        | 8.1330         | 0.9970 |
|               | EBT   | 0.8041        | 9.3600         | 0.9960 |
|               | RF    | 0.9309        | 9.1715         | 0.9962 |
|               | ANN   | 3.1310        | 7.0060         | 0.9977 |
| Barrancominas | EMP   | 6.4334        | 110.0163       | 0.4608 |
|               | MVEMP | 0.0122        | 36.8471        | 0.9366 |
|               | EGPR  | 0.5613        | 7.7538         | 0.9972 |
|               | EBT   | 0.0092        | 8.6066         | 0.9965 |
|               | RF    | 0.0568        | 9.1002         | 0.9961 |
|               | ANN   | 3.0637        | 6.9222         | 0.9977 |
| Chajal         | EMP   | 4.4464        | 78.6548        | 0.7128 |
|               | MVEMP | 0.3542        | 29.3354        | 0.9583 |
|               | EGPR  | 0.4612        | 6.5851         | 0.9979 |
|               | EBT   | 0.4481        | 7.0513         | 0.9976 |
|               | RF    | 0.3947        | 7.0558         | 0.9976 |
|               | ANN   | 2.6189        | 6.2072         | 0.9981 |
| Sipi           | EMP   | 5.5841        | 86.1744        | 0.6780 |
|               | MVEMP | 1.2220        | 30.6463        | 0.9575 |
|               | EGPR  | 0.4248        | 7.1162         | 0.9977 |
|               | EBT   | 0.6580        | 8.0377         | 0.9971 |
|               | RF    | 0.6185        | 7.8242         | 0.9972 |
|               | ANN   | 2.7263        | 6.3490         | 0.9982 |
| Puerto Merizalde| EMP  | 10.4462       | 94.6049        | 0.6154 |
|               | MVEMP | 1.1240        | 33.1178        | 0.9491 |
|               | EGPR  | 0.6048        | 7.0897         | 0.9977 |
|               | EBT   | 0.5904        | 7.5991         | 0.9973 |
|               | RF    | 0.5521        | 7.9230         | 0.9971 |
|               | ANN   | 2.8704        | 6.6222         | 0.9979 |
| Bogota        | EMP   | 5.5841        | 86.1744        | 0.6780 |
|               | MVEMP | 1.2220        | 30.6463        | 0.9573 |
|               | EGPR  | 0.4303        | 7.1792         | 0.9977 |
|               | EBT   | 0.6271        | 7.7407         | 0.9973 |
|               | RF    | 0.6464        | 7.7266         | 0.9973 |
|               | ANN   | 2.6964        | 6.3453         | 0.9981 |

Figure 4. Comparison of the statistical indicators (RMSE and $R^2$) for all sites studied.
4. Results and discussion

The first dataset contains information of 9 years, from January 2007 to December 2015, and it was used to calibrate the models. Data from January 2016 to December 2017 (two years) was used to validate the performance of the model. For each location, each one of the regression models was trained. For the set of empirical models (EMP), a 5-degree polynomial was fitted using Mt as predictor variable. For the set of multivariate empirical models (MVEMP), linear combinations of pairwise multiplication of variables were fitted. The tree-based ensemble models used the 9-year dataset and a cross validation scheme to optimize hyper-parameters such as the number of trees and minimum leaf size. Here, one hundred iterations using the Bayesian optimizer for hyper-parameter tuning was used. The characteristics of the resultant trees are summarized in Tables 3 and 4 for RF and EBT models, respectively. The ANN model was trained using the Mean Squared Error as the loss function and an Adam optimizer that uses the Nesterov momentum (Nadam). The architecture of the ANN involved 8 neurons in the input layer, 4 hidden layers with 10 neurons each, and one neuron in the output layer. Each neuron had a selu activation function and which weights were initialized using the lecun uniform kernel initializer. Furthermore, for this model, a standardization pre-processing was applied to the features used. This model can be considered as a deep neural network.

To evaluate the performance of the regression models, we used the following indicators: mean bias error (MBE), root mean square error (RMSE), and coefficient of determination ($R^2$). These indicators are defined as follows:

\[
MBE = \frac{\sum_{i=1}^{N}(y_i - x_i)}{N},
\]

\[
RMSE = \left\{ \frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N} \right\}^{\frac{1}{2}},
\]

\[
R^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(x_i - \bar{x})^2}
\]

where, $y_i$ is the $i_{th}$ predicted value, $x_i$ is the $i_{th}$ measured value, $x$ is the measured mean value, $\bar{y}$ is the predicted mean value and $N$ is the number of analyzed data points.

The methodology was applied to find a prediction model for locations Caruru, Barrancominas, Chajal, Sipi, Puerto Merizalde, and Bogotá, in which the solar $Dh$ was calculated for each site with the data obtained from the NSRDB Database from 2016 to 2017. Figure 3 shows the scatter plots of the actual measured $Dh$ values vs the predicted ones by all models at each location. A perfect model produces predictions on the unit-slope line. By a visual inspection of the results, it can be observed that MVEMP models are better than EMP models. However, ANN, EBT, EGPR and RF models provide better prediction results than both EMP and MVEMP models.

The performance indices of the six models are presented in Table 5 for each one of the locations. The best results are highlighted in bold for each site. Figure 4 shows the boxplots to provide information about the indicator’s distribution for each method. In general, models based on machine learning algorithms (EGPR, EBT, RF and ANN) provide RMSE values lower than 10 [W/m²], MBE values between 0.092 and 1.08, and ($R^2$) values close to 1. Note that, the EMP and MVMP have the worst performance. Also, EBT and RF models have a very similar distribution in the indicators. The model based on an ANN outperformed for all locations, providing a very low error and almost a perfect coefficient of determination $R^2$. Remember that this coefficient can be seen as the proportion of the variance in $Dh$ that is predictable from the extracted features. Following, we will analyze the importance of the features to provide information to the ANN to conduct the prediction process.

4.1. Feature importance

A feature importance analysis was conducted to better understand the learning process undergone by the ANN model. This was done by a permutation strategy [38], which randomly shuffles a single feature while leaving the others in place. This was done for each of the features while also calculating the RMSE score in each step. Then, we calculated the RMSE we obtained at every shuffle stage as a percentage variation from the original RMSE for each model when all original features are considered. This variation can be seen as how informative the feature is to conduct the prediction. The results are shown in Figure 5. The declination angle and the azimuth angle were the least important features to predict the $Dh$ based on...
the analysis. Furthermore, for the model applied to Barranquinas, shuf- fling these features resulted in a better performance of the model, which may suggest that removing these features could produce better results when creating new models. The most important feature to predict the Dh was the Gh for most of the models created, followed by the legal time and the hour angle. Interestingly, Mt was not as important as we expected. This can be the reason why the empirical models, which rely entirely on Mt, are not as accurate as those models that include Gh and time features to predict Dh.

5. Conclusion

A methodology was proposed to predict half-hourly diffuse horizontal irradiance in Colombia from two input variables using machine learning-based regression models and a feature engineering process based on simple angle calculations. The methodology was validated on six different sites with different geographical and climatic conditions in Colombia. Satellite measurements (NSRDB) of global and diffuse radiation were used for training, validation and testing of each regression model. We showed that only with information from horizontal global radiation (Gh), the geographic coordinate where the site is located at, and the time stamp the measurement is taken, we were able to obtain RMSE ranging from 5.86 to 9.36 [W/m²] and a coefficient of determination ranging from 0.9974 to 0.9983, showing that the methodology allows us to learn models to predict diffuse solar irradiance on horizontal surfaces with high accuracy for the six sites. Moreover, the permutation importance analysis showed that the features azimuth angle and declination angle had a low impact on the prediction of the Dh, while the legal time, Gh and hour angle features had the highest impact over the predictions.

Future investigations could be focused on the assessment of the methodology for different climatic zones (several cities of Colombia) and different estimation horizons. Other methods, such as Recurrent Neural Networks and Long-Short Term-Memory algorithms could also be tested to conduct diffuse irradiance variable forecasting.

Declarations

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data associated with this study has been deposited at Github at the accession URL: https://github.com/SmartSystems-UniAndes/Prediction_Solar_DHI_Colombia.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] G. Narvaez, L.F. Giraldo, M. Bressan, A. Pantoja, Machine learning for site-adaptation and solar radiation forecasting, Renew. Energy 167 (2021) 333–342.
[2] A.P. de Oliveira, A.J. Machado, J.F. Escobedo, J. Soares, Diurnal evolution of solar radiation at the surface in the city of siao paulo: seasonal variation and modeling, Theor. Appl. Climatol. 71 (3) (2002) 231–249.
[3] E. Paulus, R. Blaga, A simple and reliable empirical model with two predictors for estimating 1-minute diffuse fraction, Sol. Energy 180 (2019) 75–84.
[4] M. Sengupta, Y. Xie, A. Lopez, A. Habib, G. Maclaurin, J. Shelby, The national solar radiation data base (nsrdb), Renew. Sustain. Energy Rev. 89 (2018) 51–60.
[5] L. Wang, Y. Lu, L. Zou, L. Feng, J. Wei, W. Qin, Z. Niu, Prediction of diffuse solar radiation based on multiple variables in China, Renew. Sustain. Energy Rev. 103 (2019) 151–216.
[6] K. Gopinathan, A. Soler, Diffuse radiation models and monthly-average, daily, diffuse data for a wide latitude range, Energy 20 (7) (1995) 657–667.
[7] Agbadut, A. E. Gürel, Y. Binç, Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and Comparison 135 110114. Renew. Sustain. Energy Rev.
[8] L. Benali, G. Nottom, A. Fouilloy, C. Voyant, R. Dize, Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components, Renew. Energy 132 (2019) 871–884.
[9] K. Basaran, A. Ozçift, D. Kilinc, A new approach for prediction of solar radiation with using ensemble learning algorithm, Arabian J. Sci. Eng. (Springer Science & Business Media BV) 44 (8) (2019) 7159–7171.
[10] A.F. Zambrano, L.F. Giraldo, Solar irradiance forecasting models without on-site training measurements, Renew. Energy 152 (2020) 557–566.
[11] B.Y. Liu, R.C. Jordan, The interrelationship and characteristic distribution of direct, diffuse and total solar radiation, Sol. Energy 4 (3) (1980) 1–19.
[12] D. Erbs, S. Klein, J. Duffie, Estimation of the diffuse fraction for hourly, daily and monthly-average global radiation, Sol. Energy 28 (4) (1982) 293–302.
[13] A. Skartveit, J.A. Olesen, A model for the diffuse fraction of hourly global radiation, Solar Energy 38 (4) (1987) 221–224.
[14] E.L. Maxwell, A quasi-physical model for converting hourly global horizontal direct normal insolation, Tech. Rep., United States (1987) arXiv: Sponsor Org. URL, https://wwwosti.gov/serverts.parl/5987868.
[15] R. Perez, R. Seals, A. Zelenka, P. Ineichen, Climatic evaluation of models that predict hourly direct irradiance from hourly global irradiance: prospects for performance improvements, Sol. Energy 44 (2) (1990) 99–108.
[16] E.F. Abreu, P. Canhoto, M.J. Costa, Prediction of diffuse horizontal irradiance using a new climate zone model, Renew. Sustain. Energy Rev. 110 (2019) 28–42.
[17] D. Reindl, W. Beckman, J. Duffie, Diffuse fraction correlations, Sol. Energy 45 (1) (1990) 1–7.
[18] A. Skartveit, J.A. Olesen, M.E. Tuf, An hourly diffuse fraction model with correction for variability and surface albedo, Sol. Energy 63 (3) (1998) 173–183.
[19] P. Liu, X. Tong, J. Zhang, P. Meng, J. Li, J. Zhang, Estimation of half-hourly diffuse solar radiation over a mixed plantation in north China, Renew. Energy 149 (2020) 1360–1369.
[20] C. Remo, F. Petito, A. Gatto, Ann model for predicting the direct normal irradiance and the global radiation for a solar application to a residential building, J. Clean. Prod. 135 (2016) 1298–1316.
[21] J. Soares, A.P. Oliveira, M.Z. Boznan, P. Makar, J.F. Escobedo, A.J. Machado, Modeling hourly diffuse solar radiation-in the city of siao paulo using a neural-network technique, Appl. Energy 79 (2) (2004) 201–214.
[22] H.K. Elimir, Y.A. Arzam, F.I. Younes, Prediction of hourly and daily diffuse horizontal fraction using neural network, as compared to linear regression models, Energy 32 (8) (2007) 1513–1523.
[23] S. Alam, S. Kaushik, S. Garg, Assessment of diffuse solar energy under general sky condition using artificial neural network, Appl. Energy 86 (4) (2009) 554–564.
[24] N. Kaushik, R. Thomas, S. Kaushik, Artificial neural network model based on interrelationship of direct, diffuse, and global solar radiations, Sol. Energy 103 (2014) 327–342.
[25] M.A. Hassan, A. Khalil, S. Kaseb, M. Kassem, Exploring the potential of tree-based ensemble methods in solar radiation modeling, Appl. Energy 203 (2017) 897–916.
[26] H. Khorasanizadeh, K. Mohammadi, N. Goudarzi, Prediction of horizontal diffuse solar radiation using clearness index based empirical models; a case study, Int. J. Hydrogen Energy 41 (47) (2016) 21888–21898.
[27] J. Duffie, W. Beckman, Solar Engineering of Thermal Processes, Wiley, 2013. URL, https://books.google.com.co/books?id=SzIDULmGXYQC.
[28] T. Khatib, W. Elmenreich, Modeling of Photovoltaic Systems Using Matlab: Simplified green Codes, John Wiley & Sons, 2016.
[29] J.W. Spencer, Fourier series representation of the position of the sun, Search 2 (5) (1971) 172. URL, http://www.mail-archive.com/suninfo@uni-koeln.de/msg10150. html.
[30] J. Rui, H. Zhang, Q. Ren, L. Yan, Q. Guo, D. Zhang, Toc content prediction based on a combined Gaussian process regression model, Mar. Petrol. Geol. 118 (2020) 104429.
[31] P. Sollich, C.K. Williams, Understanding Gaussian process regression using the equivalent kernel, in: International Workshop on Deterministic and Statistical Methods in Machine Learning, Springer, 2004, pp. 211–228.
[32] T. Yiu, Understanding random forest, URL, https://towardsdatascience.com/understanding-random-forest-58381e606022.
[33] A. Nagpal, Decision tree ensembles- bagging and boosting, URL, https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a6bb4.
[34] H. Mark, H. Martin, D. Howard, Neural Network Toolbox™ User’s Guide, Revised for Version 11.0 (Release 2017b) Edition, The MathWorks, Inc., 2017.
[35] G. Boccaleya, Introduction to Neural Networks 892785047, 1997.
[36] M. Cerliani, Feature importance with neural network, URL, https://towardsdatascience.com/feature-importance-with-neural-network-34ebc6205743.