



**UNIVERSITAT POLITÈCNICA DE CATALUNYA  
BARCELONATECH**

---

**Escola Tècnica Superior d'Enginyeria  
de Telecomunicació de Barcelona**

**SOLAR RADIATION ESTIMATION MODELS BASED ON ARTIFICIAL  
INTELLIGENCE APPLIED TO THE PHOTOVOLTAIC ELECTRICAL  
GENERATION FOR NORTE DE SANTANDER, COLOMBIA**

**A Master's Thesis**

**Submitted to the Faculty of the**

**Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona**

**Universitat Politècnica de Catalunya**

**by**

**Mario Joaquin Illera Bustos**

**In partial fulfilment**

**of the requirements for the degree of**

**MASTER IN ELECTRONIC ENGINEERING**

**Advisor: Francisco Juan Guinjoan Gispert**

**Co-director: Sergio Basilio Sepúlveda Mora**

**Barcelona, January 2019**

## **Abstract**

The absence of direct measurements of solar radiation in many countries around the world (due to the high costs of installation and maintenance of the measuring devices) has been identified (in the Colombian case by the Energy Mining Planning Unit) as one of the main barriers for the deployment of photovoltaic systems. In this sense, estimation techniques have been developed in the literature for locations where this variable is not measured. These techniques take advantage of the correlation between the irradiance and other climatic parameters of wider distribution and easier access to construct models that forecast with high accuracy the solar potential of a specific place. Thus, the current research exposes the implementation of an indirect estimation model designed with Artificial Intelligence that uses the temperature, humidity, wind speed and sunshine duration to predict the irradiation, as a tool for sizing photovoltaic systems in Norte de Santander (Colombia) where solar radiation measurements are available just in three of the 40 municipalities in which the region is geographically divided.

# Table of Contents

<b>Chapter 1 - Introduction and objectives</b> .....	<b>8</b>
<b>Chapter 2 - Irradiation indirect estimation models: State of art</b> .....	<b>11</b>
2.1. Statistical models .....	11
2.2. Artificial Intelligence .....	11
2.3. Physical models .....	11
2.4. Empirical models .....	12
2.4.1. Sunshine-based models .....	12
2.4.2. Cloud-based models .....	12
2.4.3. Temperature-based models .....	12
2.4.4. Other meteorological parameter-based models .....	13
2.5. Hybrid models .....	13
<b>Chapter 3 - Artificial Intelligence Techniques: ANN and ANFIS</b> .....	<b>15</b>
3.1. Artificial Neural Network .....	15
3.1.1. Overview .....	15
3.1.2. Topologies .....	17
3.2. Adaptive-Network-based Fuzzy Inference System (ANFIS) .....	21
3.2.1. Overview .....	21
3.2.1 Topology .....	22
<b>Chapter 4 - Methodology</b> .....	<b>26</b>
4.1. Characteristics of the case study .....	26
4.1.1. Climatological information sources in Norte de Santander .....	27
4.1.2. Climatological information available in Norte de Santander .....	30
4.2. Development of the model .....	33
4.2.1. Pre-processing stage .....	34
4.2.2. Design stage: Structure of the network .....	47
<b>Chapter 5 - Comparison for a specific case: PV sizing</b> .....	<b>73</b>
5.1. Grid-connected PV system sizing .....	73
5.1.1. PV module selection .....	76
5.1.2. Selection of the inverter .....	76
5.2. Stand-alone PV system sizing .....	79
5.2.1. Numerical method .....	81
<b>Chapter 6 - Expansion of the models</b> .....	<b>94</b>
<b>Chapter 7 - Conclusions and recommendations</b> .....	<b>97</b>
<b>Appendices</b> .....	<b>98</b>
<b>References</b> .....	<b>105</b>

# List of figures

<b>Figure 1.</b> Distribution of studies with respect to the prediction technique.....	14
<b>Figure 2.</b> A three-layer artificial neural network .....	16
<b>Figure 3.</b> ANN of three layers.....	18
<b>Figure 4.</b> One trajectory of execution of the backpropagation algorithm. ....	19
<b>Figure 5.</b> Back propagation of the error. ....	19
<b>Figure 6.</b> Updating of the weights. ....	19
<b>Figure 7.</b> Fuzzy inference system. ....	21
<b>Figure 8.</b> Commonly fuzzy if-then rules and fuzzy reasoning mechanisms. ....	22
<b>Figure 9.</b> Basic structure of ANFIS. ....	23
<b>Figure 10.</b> Common membership functions. ....	24
<b>Figure 11.</b> Type 3 fuzzy reasoning (evaluation of the fuzzy inference).....	25
<b>Figure 12.</b> Mapping of a 2-input, type-3 ANFIS with three MF in each input. ....	25
<b>Figure 13.</b> Geographical characteristics of Norte de Santander. ....	26
<b>Figure 14.</b> Distances from Cúcuta to different municipalities. ....	27
<b>Figure 15.</b> Meteorological stations for Norte de Santander.....	33
<b>Figure 16.</b> Flowchart of the algorithm in Matlab for the first step of the preprocessing. ....	36
<b>Figure 17.</b> Flowchart of the algorithm in Matlab for the second step of the pre-processing.....	37
<b>Figure 18.</b> Standardized residual process .....	40
<b>Figure 19.</b> Flowchart of the standardized residual algorithm implemented in Matlab.....	41
<b>Figure 20.</b> Polynomial regression for a set of irradiation data.....	42
<b>Figure 21.</b> Graphical representation of the standardizing process.....	44
<b>Figure 22.</b> Flowchart for Chauvenet's criterion algorithm in Matlab. ....	45
<b>Figure 23.</b> ANN structure with the best performance for the UFPS Station.....	55
<b>Figure 24.</b> Characteristics of the training process. ....	55
<b>Figure 25.</b> Measured vs. Estimated data for hourly estimation model in the city of Cúcuta.....	58
<b>Figure 26.</b> Measured vs. Estimated data for hourly estimation model in the city of Pamplona. ....	58
<b>Figure 27.</b> Measured vs. Estimated data for hourly estimation model in the city of Herrán. ....	59
<b>Figure 28.</b> Measured vs. Estimated data for daily estimation model in the city of Cúcuta. ....	61
<b>Figure 29.</b> Measured vs. Estimated data for daily estimation model in the city of Pamplona. ....	62
<b>Figure 30.</b> Measured vs. Estimated data for daily estimation model in the city of Herrán.....	62
<b>Figure 31.</b> Measure vs. Estimated data for monthly estimation model in the city evaluated.....	63
<b>Figure 32.</b> Graphical representation of the membership functions in Matlab: Triangular.....	68
<b>Figure 33.</b> Graphical representation of the membership functions in Matlab: Trapezoidal. ....	69
<b>Figure 34.</b> Graphical representation of the membership functions in Matlab: Bell-shaped.....	69
<b>Figure 35.</b> Correlation among input variables with the output in the ANFIS structure.....	71
<b>Figure 36.</b> Relationship among input and output data from the qualitative analysis. ....	71
<b>Figure 37.</b> PSH for each one of the information sources in: a.Cúcuta. b.Pamplona. c.Herrán ..	75
<b>Figure 38.</b> Reliability map.....	80

<b>Figure 39.</b> Isoreliability curve for a specific LLP with its corresponding plot of cost indicating the optimal pair $C_{AOS} - C_{SOS}$ .....	80
<b>Figure 40.</b> Schematic of the photovoltaic system used for the numerical method. ....	81
<b>Figure 41.</b> Isoreliability curves for the cities under evaluation.....	84
<b>Figure 42.</b> Isoreliability curves comparison for the different information sources.....	86
<b>Figure 43.</b> Parameters for the phase space reconstruction. ....	89
<b>Figure 44.</b> Space phase reconstitution clustered for hourly solar radiation. ....	91
<b>Figure 45.</b> Isoreliability curves for hourly analysis. ....	92
<b>Figure 46.</b> Comparison between hourly and daily LLP method with a LLP=0,01.....	93
<b>Figure 47.</b> Comparison among available data in hourly scale for the sizing with a LLP=0,01 ...	93
<b>Figure 48.</b> Expansion of the estimation models in Norte de Santander.....	95

# List of tables

<b>Table 1.</b> Most common activation functions for an ANN.....	17
<b>Table 2.</b> Functions for the implementation of ANN training algorithms in Matlab. ....	20
<b>Table 3.</b> Common membership functions. ....	24
<b>Table 4.</b> Climatological variations among Cúcuta and some selected municipalities. ....	28
<b>Table 5.</b> Weather factors for each municipality of Norte de Santander for August 27, 2018.....	29
<b>Table 6.</b> Irradiation data by information sources and scales. ....	30
<b>Table 7.</b> Main categories of the meteorological stations of the IDEAM. ....	31
<b>Table 8.</b> Active meteorological stations for Norte de Santander. ....	32
<b>Table 9.</b> Distribution of the amount of data provided by the IDEAM. ....	34
<b>Table 10.</b> Days deleted from organized data because of the excess of missing data. ....	37
<b>Table 11.</b> Percentage of missing values: a. Interpolated. b. Extrapolated. ....	38
<b>Table 12.</b> Total amount of data points for the outlier detection algorithms.....	38
<b>Table 13.</b> Example data for the application of the standardized residual criterion. ....	40
<b>Table 14.</b> Standardized residuals for data in table 13.....	40
<b>Table 15.</b> Fit indicators for each variable under evaluation. ....	41
<b>Table 16.</b> Percentage of modified data by the analysis of standardized residuals. ....	42
<b>Table 17.</b> Percentage of modified data by the analysis with Chauvenet's criterion. ....	44
<b>Table 18.</b> Final amount of data points for the normalization process.....	46
<b>Table 19.</b> Training parameters. ....	50
<b>Table 20.</b> Performance indicators for each process of construction of the model in terms of the number of neurons in the hidden layer. ....	50
<b>Table 21.</b> Performance comparison among learning algorithms for a feed-forward back-propagation topology of 3 layers.....	51
<b>Table 22.</b> Comparison in terms of performance for different combination of the activation functions. ....	52
<b>Table 23.</b> Performance of three topologies for the best behavior of the design parameters evaluated.....	52
<b>Table 24.</b> Final characteristics of the implemented ANN.....	52
<b>Table 25.</b> Correlation among climate variables and solar radiation. ....	53
<b>Table 26.</b> Performance of the ANN for different combination of the input variables. ....	53
<b>Table 27.</b> Allocation of weights by the training algorithm.....	54
<b>Table 28.</b> Weights between hidden and output layer for UFPS station.....	54
<b>Table 29.</b> Weights from the training of the ANN. ....	56
<b>Table 30.</b> Weights between hidden and output layer for the ANN developed.....	56
<b>Table 31.</b> Error indicators for ANN model in each one of the evaluated cities.....	57
<b>Table 32.</b> Error indicators for daily and monthly estimation models.....	60
<b>Table 33.</b> Error indicator for several models in the literature.....	66
<b>Table 34.</b> Parameters for the implemented membership functions.....	68

<b>Table 35.</b> Performance of the ANFIS implementations. ....	70
<b>Table 36.</b> Error indicators for the develop ANFIS models for the evaluated cities in the estimation ranges: Hourly, daily and monthly. ....	72
<b>Table 37.</b> Comparison between the ANN and ANFIS results in terms of the MAPE indicator. ...	72
<b>Table 38.</b> Results of the PV+Test for the top 9 PV panels. ....	76
<b>Table 39.</b> Characteristics of the selected PV module. ....	77
<b>Table 40.</b> Technical data for the selected inverters. ....	77
<b>Table 41.</b> PV system sizing for a residential application.....	78
<b>Table 42.</b> Limits to classify a month and a day with low irradiation. ....	83
<b>Table 43.</b> Optimal pairs $C_A - C_S$ for different LLP values in daily scale analysis.....	84
<b>Table 44.</b> PV system sizes for different information sources in terms of the autonomy days.....	87
<b>Table 45.</b> Optimal pairs $C_A - C_S$ for different LLP values in daily scale analysis.....	92
<b>Table 46.</b> Main geographical characteristics and climate conditions for the cities evaluated.....	94
<b>Table 47.</b> Comparison of the error indicators for the Herrán's model in other cities.....	95
<b>Table 48.</b> Stations recommended for the construction of irradiation estimation models with application in Norte de Santander. ....	96

# Chapter 1

## Introduction and Objectives

Solar radiation is the energy emitted by the Sun, which propagates in all directions through space as electromagnetic waves, generated due to hydrogen reactions in the core of the Sun through nuclear fusion. This energy is the engine that determines the dynamics of atmospheric processes and climate on Earth; of the 64 million  $W/m^2$  produced on average by the surface of the Sun, the Earth intercepts only  $1.367 W/m^2$ , which are distributed in a greater proportion in the tropical zones (countries near to the Equator Line). This is because the ultraviolet rays affect these regions more directly than those with middle latitudes, and also because their ozone levels are lower (close to 240 UD -Dobson units, being  $UD = 2,69 \times 10^{16}$  molecules/cm<sup>2</sup>) than the average ones in the northern and southern areas of the planet (taking into account that ozone is a factor inversely proportional to the incidence of solar radiation on the surface of the Earth [1]).

Colombia is located in the tropical zones close to the Equator with a wide variety of climates highly dependent on the altitude with respect to sea level; a preliminary analysis determined that Colombia has a significant solar potential with average irradiance values higher than the world average (of up  $6,237 kW/m^2$  in comparison with the world average of  $3,9 kW/m^2$ ) and the average of countries in the northern and southern hemisphere [2].

In this way, Colombia become an ideal place for the implementation of photovoltaic (PV) systems by the high solar radiation levels of its geographical zone.

Despite of the above, the Energy Mining Planning Unit of Colombia (UPME, by its acronym in Spanish) declared in its balance for 2017 that the photovoltaic capacity installed in the country is close to 0,06 % of the total capacity of electricity generation [3], condition that in the last years has been a center of discussion bearing in mind that solar energy is one of the alternatives with greater possibilities of success in the country, and considering the need of diversifying the energy matrix of the nation. The later in response to the high dependence from the hydrological resource in Colombia (approximately 70 % of participation) which has a wide number of advantages as clean energy production, renewable resources, safety, but also has a hard sensibility to the changing weather conditions of the region where environmental phenomenon as *El Niño* or *La Niña* are very common.

The low participation of PV systems is attributed by the UPME among several causes, to the lack of technical specifications for its application and the limited information about the energy potential that exists in each zone of the country. The UPME considers these situations as two of the identified and prioritized barriers for the growth of the photovoltaic solar energy in Colombia [2], scenario that has also been identified in other places [4].

The previous argument can be based on the fact that the Institute of Hydrology, Meteorology and Environmental (IDEAM, by its acronym in Spanish), entity in charge of the management of the weather information in Colombia, has 4.451 meteorological stations throughout the country [5] but only 83 of these record global solar radiation data [6]; this scant coverage does not allow to determine the specific and detailed solar potential in the Colombian territory and it suggests that the average irradiance values supplied by the IDEAM could be inaccurate for PV design (causing oversizing or under-sizing of the system and erroneous periods of return of the investment, which leads to a loss of the reliability of the users on this type of technologies) in regions that are distant from those 83 locations where the solar radiation is measured.



Therefore, one of the key points in the development of the PV technology in Colombia depends on the accurate quantification of the solar resource in those places where irradiation data are absent, and that in general, represent the rural areas of the country.

The aforementioned scenario is not only present in Colombia, but it is also common in developing countries [7] and even in some developed countries such as China [8]. The measurement of solar radiation data is generally available just in some specific areas due to the acquisition and maintenance costs of the solar radiation measuring devices in comparison to those that measure other variables. For example, the ratio between stations observing solar radiation and those observing ambient temperature is lower than 1:100 in America [9]. Due to this limitation, variables such as humidity, temperature, wind speed, among others, are much easier to access than the solar radiation.

Considering that several studies have shown the correlation between the solar radiation and these climatic variables of greater availability [10] [11] [12] [13], different modelling techniques have been developed to estimate the solar radiation from them with high accuracy. This allows the expansion of reliable irradiation profiles to locations where this variable is not measured but where those correlated variables are available, improving the design process of PV systems; thereby, the applications based on solar energy increase [8] [14] [15] and their penetration in a specific place also makes it [16].

In the literature, these structures that use climatological parameters as input to generate a corresponding irradiation value are called *indirect estimation models*, since the model forecasts the solar radiation from other variable and then, using a PV performance model of the plant or a commercial PV simulation software, such as TRNSYS, PVFORM, or HOMER, calculates the PV power generation. In a similar way, the *direct* version can be also found, but this handles historical data samples, such as PV power output and meteorological data associated to predict directly the energy produced by the system [16] [17]. As it can be intuited, the direct estimation would be more precise because it would take into account the contribution of the behavior of the panel to record output data from the system implemented; however, it also becomes a disadvantage due to the need of a physical prototype for its construction (which could be unfeasible as a previous investment is required to determine if systems in the location under analysis are profitable or not; additionally, its realization requires a considerable amount of time taking all the samples for the evaluation).

Although nowadays, there is a wide variety of satellite information sources that provide irradiation data, the results of these models constructed with ground-measurements in many investigations have demonstrated to have a higher performance with minimum complexity and computational cost.

Thus, based on the necessity of obtaining more precise and detailed information of the solar resource for its application in different fields, and the solution offered by the estimation models, the government of Colombia has implemented a set of initiatives aimed to projects that use its extensive climatological stations network to predict the irradiation in all regions of the country. One of these initiatives is named PERS (Sustainable Rural Electrification Plan), whose goal is to develop indirect estimation models to generate and test new solar radiation data that supports and complements improves the information supplied today by the IDEAM.

Norte de Santander is one of the 32 departments in which Colombia is geographically divided, and it was selected for the PERS initiative and it is also the focus of this work. PERS in this region resulted with the implementation of three empirical models based on the relationship between solar radiation and the sunshine hours; although these models were constructed with data of three specific locations, PERS researchers indicated that it was possible to implement the models

in different cities and rural zones of the Department. Details about this will be explained in next sections.

Hence, considering all of the above, the goal of this research is to develop indirect estimation models (since according to the limitations for a physical implementation of the system respect to the time to develop this project, a direct model could not be performed) based on Artificial Intelligence (AI) for Norte de Santander, in order to extend the results reached by PERS in this Department (taking into account that AI models have shown better results than empirical models in the literature) and to contribute to the efforts of the government of Colombia for incrementing the tools and information for the sizing and implementation of PV systems, mainly in the rural zones.

Consequently, the general approach of the present investigation is to define the advantages that indirect estimation models using AI could offer in the solar radiation database in the Department of Norte de Santander; and the specific objectives are detailed as follows:

- Analyze the meteorological information available for Norte de Santander and determine its correlation with the solar radiation.
- Construct indirect estimation models for irradiation data based on Artificial Intelligence in the zones of the Department where these data are measured.
- Compare the performance of the AI models with similar works in the literature.
- Compare the results from AI models with other information sources available for the region in terms of PV sizing.
- Analyze the scope of the AI models to zones of the Department where irradiation data are not available.

In reference to these objectives, the outline of the document is: After the introduction, Chapter 2 defines the state of the art of the indirect estimation models applied to irradiation profiles. In Chapter 3, a brief review about the AI techniques used in the development of these models is displayed. Chapter 4 shows the design methodology of the models for the zones under evaluation; in this section a short overview with the characteristics of the Department of Norte de Santander is also depicted to contextualize the reader with the feasibility of this type of research in that region. A comparison of results is performed in Chapter 5, and the applicability of this research in Norte de Santander is exposed in Chapter 6. Finally, the conclusions of the work are presented in Chapter 7.

The choice of Norte de Santander as a case study site stems from the fact that the master studies of the author were funded by the government of this Department for the realization of this research.

# Chapter 2

## Irradiation indirect estimation models: State of the art

In the literature, many indirect estimation models applied to the identification of the irradiance profiles have been developed. As it was previously mentioned, these models aim to forecast the solar potential in places where irradiation data based on ground-measurements are not available. Their objective is to predict with the highest accuracy the actual irradiation to perform appropriate sizing and economic analysis of a PV system.

Among the main theoretical indirect forecasting methodologies are: statistical approach, artificial intelligence techniques, physical models (mainly based on numerical weather prediction, NWP), and hybrid strategies [18]. Additional to this, others as probabilistic and empirical forecasting methods, and weighted Gaussian process (GP) regression have become a recent research focus [19]. A brief explanation of the previous methodologies with some investigations is presented below.

### 2.1. *Statistical models:*

Statistical approaches have been widely used in time series forecasting. In general, these are based on historical data. The predictor constructs a statistical relationship between the variables used as inputs for the model and the variable to be predicted [18] [20]. The statistical models most used for the prediction of solar radiation are ARMA and ARIMA [21] [22] [23], but other methods as Persistence [24] [25] and ARMAX [26] are also used.

### 2.2. *Artificial Intelligence:*

AI techniques are being used in various fields, including forecasting, pattern recognition, control, optimization, and so on. Due to the high learning and regression capabilities, AI techniques have been widely employed for modeling and prediction of solar energy [18]. Among the most notable models in this field are: Artificial neuronal network [4] [27] [28] [29] [30] [31] [32], fuzzy logic [33] [34], support vector machine [35] [36] [37] and genetic algorithms [38].

### 2.3. *Physical models:*

Physical models consist of the set of mathematical equations that describe the physical state and dynamic motion of the atmosphere. These are designed with the PV power plant characteristics, such as location, different meteorological variables, and orientation historical data. In this type of models two methodologies are predominant: Sky Image-Based Models and Numerical Weather Prediction (NWP) Based Models. The first methodology is developed with satellite images and with ground-based sky image approaches; and the second one, with numerical dynamic modeling of the climatological conditions.

In reference to the sky image methodology, the satellite images are used to detect the motion of cloud using motion vector fields, which allow to identify the clarity of the trajectory of the solar radiation and therefore, the amount of energy that can be captured for a system in ground. These images are obtained from satellite stations such as: Meteosat Satellite Network (Europe, Central Asia and Africa), Geostationary Operational Environment Satellite (America) [18], Geostationary Meteorological Satellite GMS - Himawari and MTSAT satellites (East and Southeast Asia and Oceania) [39]. The information given by these stations can be found in several databases as:

Satel-Light, Solar Data (SoDa), NASA Surface Meteorology and Solar Energy (SMSE) and the Australian Bureau of Meteorology (ABM) [39]. The methods and algorithms for processing the captured images and then, to perform the solar estimation process vary from one database to another. Heliosat is one of the most used methods in several research studies based on data from the Meteosat satellites. In this field different studies about the spatial-temporal variations of solar radiation around the world have been performed [40] [41] [42].

In contrast to the satellite image-based method, ground-based sky images can provide a much higher spatial and temporal resolution for solar forecasts, on the basis of a total sky imager (TSI); this is an equipment which has a hemispherical mirror with a downward position pointing towards a couple-charged device (CCD) camera located above it. The mirror is equipped with a Sun tracking shadow band that shields the optical sensor from the consequences of solar reflection.

Despite of satellite analysis has increased its participation in recent years with desirable results in the solar prediction field, it has been found that the results estimated of solar radiation using remote sensing methods may not be as good as empirical and artificial intelligence methods, although the satellite ones can provide a wider spatial distribution in regional or global scales [43]. One of the reasons for this is argued in the fact that satellite models require much more elements to perform an accurate estimation which are not available in most cases; some of these elements can be complex atmospheric models, ground-measurements such as atmospheric turbidity, and the ability to differentiate between cloud- and ground-reflected radiation during the day [9] [44].

The decision between the advantages of a wider spatial distribution and the loss of accuracy in the estimation is still not clear; a comparative analysis between satellite and ground-based measurements methods in [45] found that factors as landform, latitude and weather station density (i.e., the number of stations by km<sup>2</sup>) can increase or decrease the accuracy of each method under different application cases. At the end, the conclusion is that the selection is not between satellite-derived and ground-based data, but between which satellite model and ground measurement model can be used.

#### 2.4. Empirical models:

These models have been developed to estimate daily or monthly global solar radiation by means of correlations with the more readily available meteorological data at a majority of weather stations [46]. Selecting an appropriate model from various existing design strategies depends on the available data of the place and the accuracy offered by the method to develop the model. Despite of the many empirical correlations reported by different authors, these models can be mainly classified into the following four categories.

2.4.1. *Sunshine-based models*: The most commonly used parameter for estimating global solar radiation is sunshine duration, since it is a variable directly proportional to the changes of the solar resource. Sunshine duration can be easily and reliably measured from devices called Heliographs, so that the data are widely available at the weather stations. Among the models that use the sunshine as input parameter are: Angstrom–Prescott model and Glowerand McCulloch model.

2.4.2. *Cloud-based models*: The cloud data are detected routinely by meteorological satellites, so global solar radiation can be estimated from observations of various cloud layer amounts and cloud types. These models differ from the physical ones in the complexity of the analysis, due to that physical approaches do not use only the cloud information for the prediction. Black model and Paltridge model are the most representative in this field.

2.4.3. *Temperature-based models*: The temperature-based models assume that the difference between the maximum and minimum temperature is directly related to the fraction of

extraterrestrial radiation received at the ground level. Hargreaves model and Bristow and Campbell model are widely used in forecasting of solar radiation from temperature.

*2.4.4 Other meteorological parameter-based models:* Many researchers have tried to use various available meteorological parameters such as precipitation, relative humidity, dew point temperature, soil temperature, evaporation and pressure to predict the amount of global solar radiation. Some works are shown as follows:

- Swartman and Ogunlade model: Use the relative humidity as correlation variable to determine the solar radiation.
- Garg and Garg model: It includes a precipitation factor for the prediction.

In [46] and [47] an extensive review of different models (with the highest accuracy in their respective works) in each category is presented for easing the choice of a model according to the characteristics and available data of the implementation place. It is worth to note that previous investigations are only a guideline to select the best type of correlation, since the coefficients of the equations resulted from empirical analysis depends on the climate conditions of the specific location where the research was carried out.

### *2.5. Hybrid models:*

To further optimize the accuracy in the prediction of solar radiation, several authors have proposed combining different models and techniques [48] [49] [50]. This method has been widely used and it presents combinations of, for example, satellite images as input parameters to an artificial neural network or even the participation of three different strategies in order to reach the highest accuracy with respect to individual approaches.

In this sense, there is a large list of methods or techniques for the estimation of the solar radiation; each one of them with its advantages and disadvantages according to the location, the availability of the input parameters and the desired spatial-temporal forecasting, which makes it complicated to determine a global selection pattern when a model is needed. Considering this, several review papers which group and compare a large amount of models [8] [9] [16] [17] [51] were analyzed to determine which of them could generate more benefits for the research. After this evaluation, the AI technique based on Artificial Neural Networks (ANN) was selected taken into account in the following reasons:

1. Wide literature: The authors in [17] compared about 84 research studies and found that models structured with ANNs are the most used machine learning technique in solar power forecasting; this is because, in average, ANNs presented the best results in terms of accuracy regarding typical performance metrics in this field as root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), coefficient of determination ( $R^2$ ) and mean bias error (MBE).

Figure 1 describes a distribution of studies with respect to the technique used, where it can be observed that the highest participation is for ANN models with a 24 % of the total. It should be noted in figure 1 that ANN (and other artificial intelligence techniques) appears inside of the statistical group. This is because in [17] AI and statistical methods were categorized as a single group, which can be different in other research since this classification changes slightly from one investigation to another. In a similar way, a comparative review of the main estimation methods is presented in [51] with a total of 74 research studies, and in [16] where can be also observed that ANN models outperform the results of traditional methods.

2. High accuracy with respect to other models: As it was mentioned in the later point, there is a wide participation of ANN models in the solar estimation processes, due to the good results in

terms of the accuracy achieved in several works; in reference to [51], several performance parameters according to the information provided by the researchers, allowed to establish criteria to compare the models. ANN demonstrates in average the lowest errors; for instance, in terms of the RMSE, values as low as 0,2 % (Ref. 27 in [51]), 3,38 % (Ref. 30 in [51]), 5,19 % (Ref. 32 in [51]) and 8,81 % (Ref. 31 in [51]) were obtained, while methods as ELM, ARIMA, Sky images and Satellite images obtained RMSE values of 13,83 % (Ref. 39 in [51]), 29,73 % (Ref. 57 in [51]), 11,17 % (Ref. 79 in [51]) and 15,47 % (Ref. 82 in [51]), respectively.

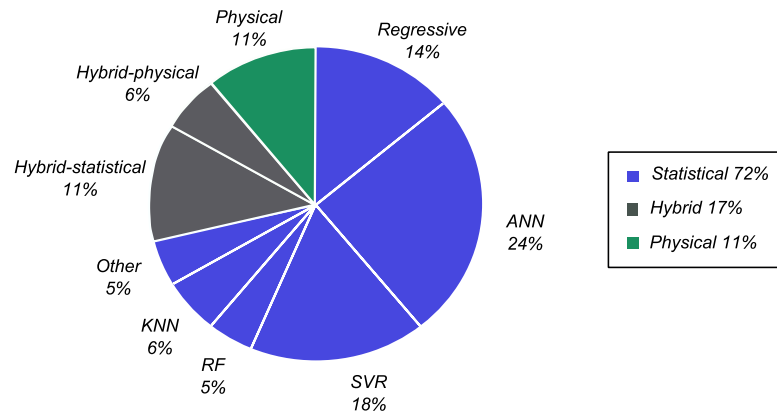


Figure 1. Distribution of studies with respect to the prediction technique used. Sample set: studies listed in [17].

- Advantages in relation to the available data: The capability of using several input parameters in ANN models represents an advantage (unlike many empirical and statistical models which use only one variable), considering that a wide variety of research findings have proved that this characteristic improves the precision in the output prediction [8] [16]. Thus, because different climatic variables data are available in Colombia (as it will be detailed later), this advantage can be exploited in the development of the model. Additionally, ANN techniques do not require a linear relationship between the input and the output of the model, being ideal for nonlinear correlations as those presented by available variables such as the humidity and wind speed with the solar radiation; this is an interesting approach since statistical model as auto-regressive and regressive-based models do not support it [16].
- Adaptability to other models: Hybrid models have begun to be popular in the estimation of solar radiation due to the increase of the efficiency by taking the individual advantages of each model and deleting its deficiencies in a cluster [16] [51]. For this reason, in order to analyze the possible adaptation of another model (which increases the accuracy as in [51] where the application of the a hybrid method between neural network and fuzzy logic, called Adaptive-Network-based Fuzzy Inference System (ANFIS), improved the accuracy in 95 % from the standard fuzzy logic), ANN is selected since in most cases, it is the common cooperador in hybrid models [16].
- Previous investigations: Since the Colombian government has performed some studies in the estimation of solar radiation (e.g. PERS project) and some solar radiation satellite databases can be accessed, empirical and physical (satellite images) models are not taken into account in the present work, leaving thus the ANN models as an alternative not yet explored with possible better results.

Thus, since the ANN model is selected as the best option for the models, the next chapter is focused on the analysis of its characteristics for its application in the chapter 4 (design and methodology). Additionally, ANFIS model is included as a hybrid method of the ANN model considering the general improvements that the combination of methods represent and the specific benefits mentioned in the previous point (Adaptability to other models).

# Chapter 3

## Artificial Intelligence Techniques: ANN and ANFIS

As it was mentioned at the end of the previous chapter, AI techniques based on ANN models are the choice for the development of the indirect estimation model of this research, for the Department of Norte de Santander. However, since hybrid models are shown as a way to improve the accuracy of the predictions, and in relation to the point 4 of the selection reasons of the ANN models in chapter 2, an ANFIS model is included in the current chapter as an additional element in the search of the best model. Thus, both AI techniques are detailed briefly below to make the reader familiar with upcoming sections.

### 3.1. Artificial Neural Network:

#### 3.1.1. Overview:

ANN is defined as non-linear map systems whose structure is based on principles observed in the nervous systems of humans and animals. It consists of a large number of simple processors linked by connections with weights. The processing units are called *neurons*. Each unit receives inputs from other nodes and generates a simple output that depends on the available local information, stored internally or arriving through the connections with the weights [52]. Thus, each process unit has a simple task: it receives input(s) from other units or external sources and processes the information to obtain an output that propagates to other units. A network can have an arbitrary structure, but the layers containing these structures are defined according to their location in the topology of the neural network. The external inputs are applied in the first layer, and the outputs are considered the last layer. Internal layers that are not observed as inputs or outputs are called *hidden layers*. By convention, entries are not considered as a layer because they do not perform processing.

The total input  $u$  of a unit  $k$  is the sum of the weights of the inputs connected, plus a bias  $\theta$  which operates as an external input for each unit:

$$u = \sum_j \omega_j x_j + \theta$$

where:

$\omega_j$  = weight of the connections

$x_j$  = output signals from other nodes or external inputs

If the weight  $\omega_j$  is positive, it represents an *excitation* and if the weight is negative, it is considered an *inhibition* of the input.

Figure 2a shows a basic structure of an ANN with three layers, where the first layer is the input layer ( $i$ ) which receives input information, the second layer is the hidden layer ( $j$ ) that can be constituted by several layers (each one with similar or different structure) and determines the processing of the input information, and finally the third layer known as the output layer ( $k$ ) where the results are received and analyzed to provide the output. Each layer is interconnected by its corresponding weight  $\omega_{ij}$  and  $\omega_{jk}$ , and every unit sums its inputs, adds the bias or threshold term

and a linear or nonlinear function transforms the sum to produce an output. This transformation is called the *activation function* of the node, which is frequently a linear function in the output layer [53].

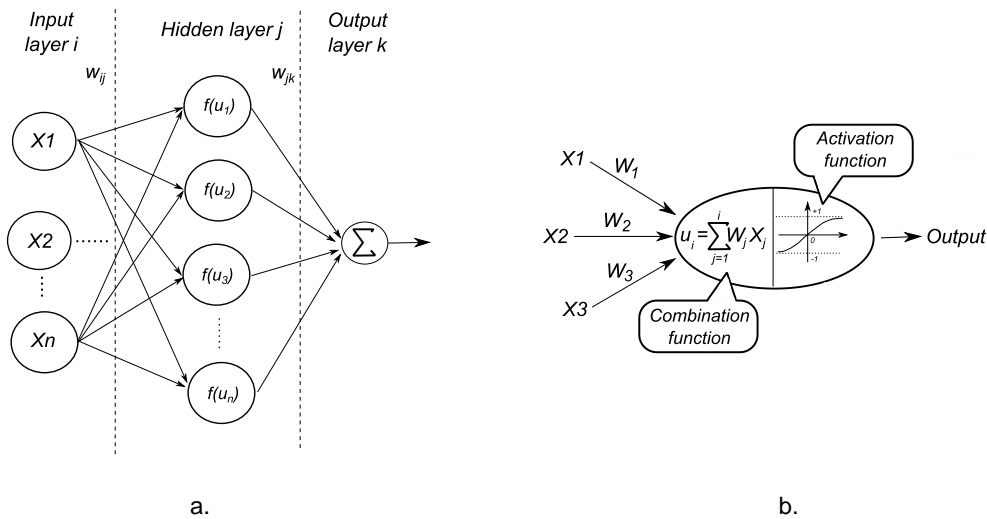


Figure 2. A three-layer artificial neural network: a. Structure. b. Mathematical model of an ANN neuron.

Based on the above, each neuron is composed by two parts: A first part in charge of the sum of the inputs called *combination function*, and a second part defined by the activation function. The most common activation functions are described in table 1 [16] [52]. The functions with bipolar operation also have their unipolar version.

Activation function	Formula	Graphical representation
Bipolar step function	$f(u) = \begin{cases} -1 & \text{if } u < 0 \\ 0 & \text{if } u = 0 \\ 1 & \text{if } u > 0 \end{cases}$	
Linear	$f(u) = u$	
Bipolar linear function	$f(u) = \begin{cases} -1 & \text{if } u < -c \\ 1 & \text{if } u > c \\ ax & \text{otherwise} \end{cases}$	
Sigmoid	$f(u) = \frac{1}{1 + e^{-u}}$	



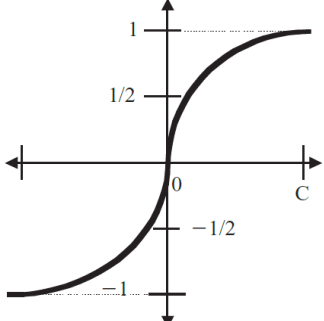
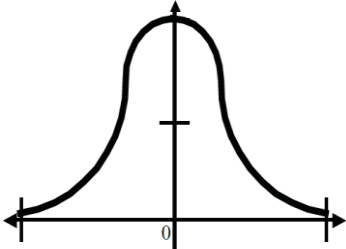
Hyperbolic tangent sigmoid	$f(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$	
Gaussian radial basis	$f(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$	

Table 1. Most common activation functions for an ANN.

In Matlab, some activation functions for the training process in ANNs are the hyperbolic tangent function (*tansig*), the sigmoid function (*logsig*), the hard-limit transfer or step function (*hardlim*) and the linear function (*purelin*), which facilitates the approximations that are required between inputs and outputs.

### 3.1.2. Topologies:

Two of the most used topologies are presented, according to the differences in the way in which the connections are made:

1. Forward propagation networks (feed-forward): the flow of information from the inputs to the outputs is exclusively forward, extending through multiple layers of units, but there is no feedback connection.
2. Recurrent networks: they contain feedback connections, which can result in a process of evolution towards a stable state in which there are no changes in the activation state of the neurons.

ANNs can also be categorized according to the number of layers (mono-layer or multi-layer) or the type of data that they receive (analog or discrete), but the information flow or connection structures is the most representative way for its classification.

The process of configuration of a neural network is called *training* or *learning* so that the inputs produce the desired outputs through the strengthening of the connections or weights. Each topology uses one or a combination of training algorithms to reach this objective. One way to accomplish this is from the establishment of previously known weights, and another method involves the use of feedback techniques and learning patterns that change the weights until appropriate values are found. In addition, learning can be divided into *supervised* or *associative* and *not supervised* or *self-organized*. In the first case, inputs that correspond to certain outputs are introduced, either by an external agent or by the same system. In the second one, the training focuses on finding statistical characteristics between groupings of patterns in the entries [52]. Part of the literature mentions a third learning strategy known as *Reinforcement Learning* [28]; this strategy lies somewhere in between the previous two types of learning. It is used in cases where an exact output of the problem modeled is unknown.

One type of rule that is used for training by adjusting the weights is the *Hebbiana*, proposed by Hebb in 1949. Several proposed variants of Hebbiana have appeared over time. If two units  $j$  and  $k$  are active at the same time, the connection between the two must be strengthened by modifying the weight:

$$\Delta\omega_{jk} = \gamma y_j y_k$$

where  $\gamma$  is a constant of positive proportionality that represents the learning rate, and  $y_j$  together with  $y_k$  represent the output of the unit.

Another rule commonly used involves adjusting the weights through the difference between the current and the desired activation; it is known as the *Delta Rule*:

$$\Delta\omega_{jk} = \gamma y_j (d_k - y_k)$$

where  $d_k$  is the desired activation.

Within supervised learning three major algorithms can be distinguished to carry out the training: learning by error correction, which takes the difference between the actual and estimated output and sets the weights to reduce the error. Learning by reinforcement, is a slower training technique than the previous one; it adjusts the weights based on probabilities because unlike the previous case, this learning strategy only evaluates if the output is correct or incorrect with respect to the input; and the last one is a stochastic learning algorithm which consists of randomly modifying the connection weights within the network and evaluating if it meets the requirements or not. If the answer improves, the algorithm modifies the weights, and if not, generates new random weights, until the expected response is reached.

Inside of the first category of the supervised learning algorithms, i.e., learning by error correction, is found the *back-propagation algorithm*. It is one of the most popular ones since it has an optimization method that defines the error gradient and minimizes it with respect to the parameters of the neural network [52]. A brief explanation about it is displayed below.

Figure 3 shows a three-layered network with two inputs and one output; each neuron consists of two units, where the first one adds the products of the inputs with their respective weights, and the second unit contains the activation function.

To train the neural network it is necessary to use a set of data, which consists of input signals  $x_1$  and  $x_2$  assigned with a corresponding objective (desired outputs) called  $y$ . Training is an iterative process, and each cycle or iteration is known as *epoch*. In each epoch the weights of the nodes are modified using a new set of data for the training. Figure 4 depicts a possible trajectory from the inputs to the output of the ANN; the path from the inputs to the first layer is highlighted in figure 4a, and figure 4b displays the trajectory from the first to the second layer with the corresponding outputs of each one of its involved neurons. The weight  $\omega_{jk}$  indicates the connection of the output of the neuron in the layer  $j$  with the input of the neuron in the layer  $k$ .

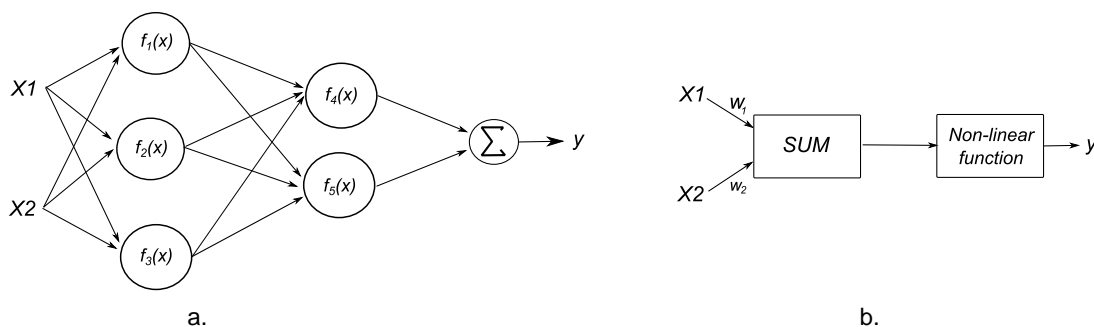


Figure 3. ANN of three layers: a. Structure of the network. b. Structure of each neuron.

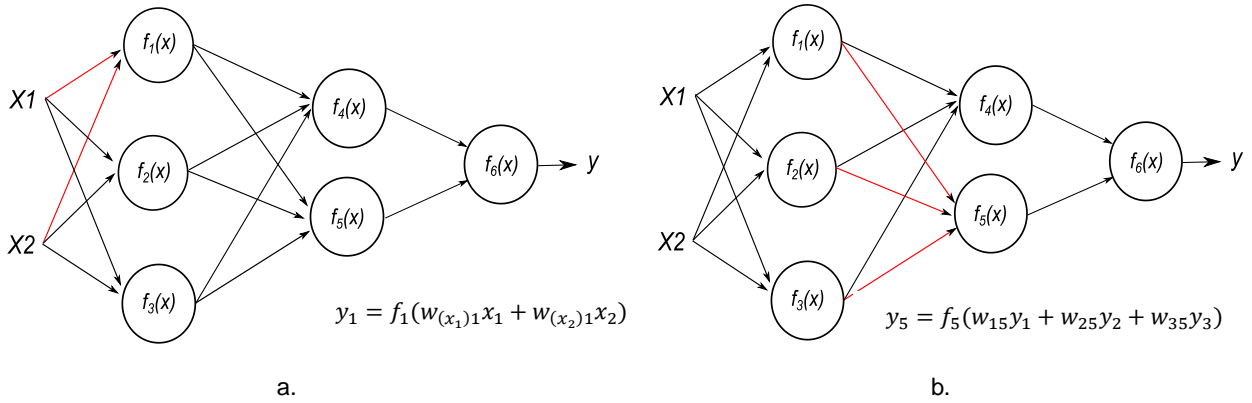


Figure 4. One trajectory of execution of the backpropagation algorithm: a. From inputs to the first layer. b. From the first layer to the second one.

After executing the route in figure 4, the output of the network is compared with the desired target value. The difference is called signal error ( $\delta$ ). It is impossible to know the error in the neurons of the internal layers directly, because the output values of these neurons are unknown. The backpropagation algorithm propagates the signal error back to all neurons, whose output was the input of the last neuron, as shown in figure 5a. Subsequently, the error is propagated to the neurons of previous layers, considering the weights of the connections, as shown in figure 5b.

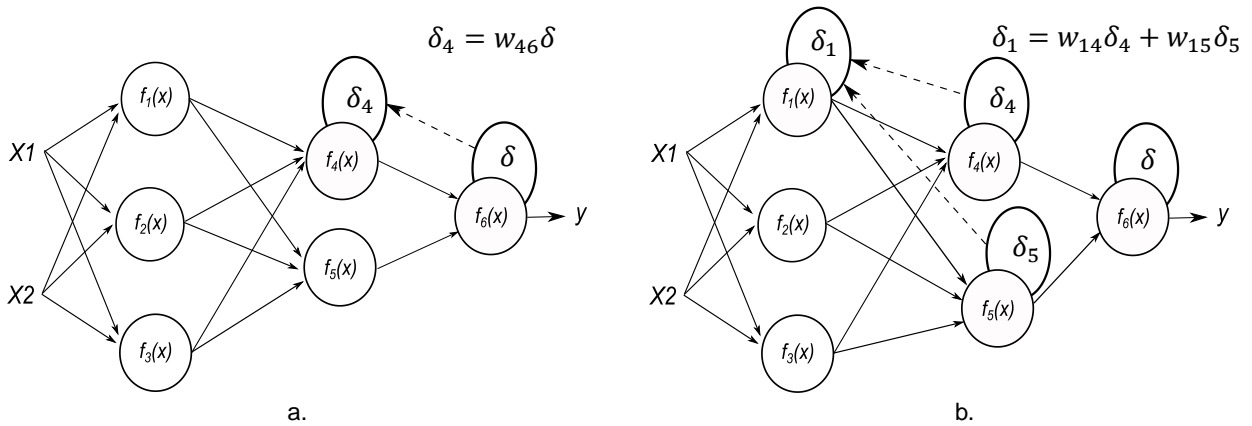


Figure 5. Back propagation of the error: a. From the last layer to the second one. b. From the second layer to the first one.

Finally, when the error is calculated for each neuron, the input weights can be modified according to the expressions in figure 6 to start a new iteration. The error is minimized across many training cycles or iterations until reaching the desired level of accuracy.

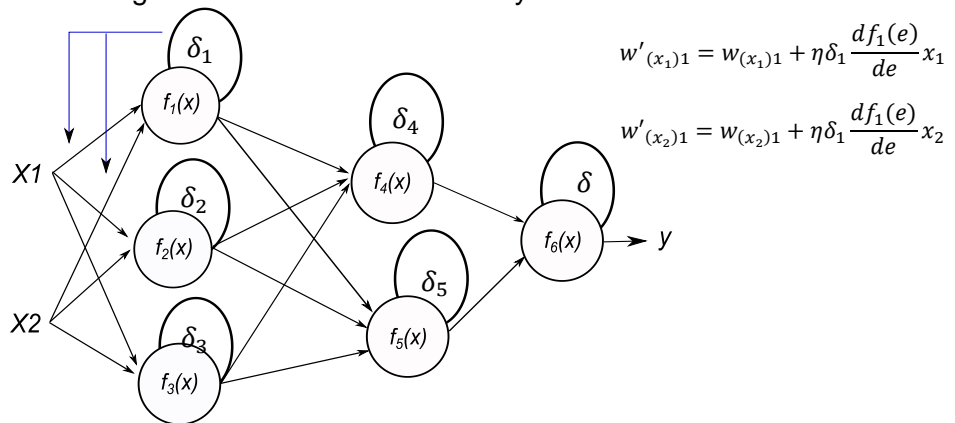


Figure 6. Updating of the weights.

The coefficients affect the learning speed and can be selected by different methods. One of them implies that at the beginning of the training process a large value is chosen, which gradually decreases as the process progresses. Another method starts with small parameters that increase as the process progresses and again decreases in the final stage. Starting the process with a small parameter allows the establishment of the signs of the weights faster and for this, it is the recommended option. It is common that the back-propagation algorithm is mixed with other algorithms to improve its performance in the development of the model. Among the most remarkable training algorithms are [53]:

- Gradient Descent (Gradient Descent back-propagation algorithm, Gradient Descent with Momentum, Resilience back-propagation).
- Conjugate Gradient algorithms (Scaled conjugate Gradient, Conjugate Gradient back-propagation with Fletcher-Reeves Updates, Conjugate Gradient back-propagation with Polak-Riebre Updates).
- Quasi-Newton algorithms (Broyden-Fletcher Goldfarb Shanno, Levenberg–Marquardt back-propagation).

The corresponding functions for the implementation of these algorithms in Matlab are deployed in table 2.

Algorithm	Function
Gradient Descent back-propagation algorithm	<i>traingd</i>
Gradient Descent with Momentum	<i>traingdm</i>
Resilience back-propagation	<i>trainrp</i>
Scaled conjugate Gradient	<i>trainscg</i>
Conjugate Gradient back-propagation with Fletcher-Reeves Updates	<i>traincgf</i>
Conjugate Gradient back-propagation with Polak-Riebre Updates	<i>traincgp</i>
Broyden-Fletcher Goldfarb Shanno	<i>trainbfg</i>
Levenberg–Marquardt back-propagation	<i>trainlm</i>

Table 2. Functions for the implementation of ANN training algorithms in Matlab.

The fact whereby the back-propagation algorithm is explained in detail in the present work, is because the Levenberg-Marquardt (LM) back propagation is one of the most used algorithm for estimation and forecasting of solar radiation in the literature [4] [29] [53] [54]. The reason for its common use is because it is considered to be one of the fastest and most accurate algorithms since it combines the speed of the Newton algorithm with the stability of the steepest descent method to reduce tasks and the sums of the squares of the new non-linear tasks. The LM algorithm uses Newton's method to calculate Jacobian matrices without computing the hessian matrices. This makes the LM algorithm to have a faster convergence with minimal error. Therefore, this is the algorithm used as starting point in the development of the ANN in the next sections.

Thus, considering the concept of the back-propagation, and the classification of the ANN topologies according to the information flow, several architectures are detached as follows [54]:

- Feed forward Network (FFN)
- Feed-Forward Back Propagation Network (FFB)
- Cascade-Forward Back Propagation Network (CFB)
- Elman Back Propagation Network (ELM)

- Radial Basis Neural Network (RB)
- Probabilistic Neural Network (PNN)
- Custom Network

Being the multilayer FFB network with LM as optimization algorithm the combination more frequently implemented in the construction of the estimation models for solar radiation [4] [16] [29] [53].

In this way, the design of the ANN for this work takes as starting parameters those with better results in the literature. By varying different components (number of neurons, number of hidden layers, activation function, optimization algorithms and topologies), the use of those starting parameters will be verified, otherwise a new implementation criterion will be established.

### 3.2. Adaptive-Network-based Fuzzy Inference System (ANFIS):

#### 3.2.1. Overview:

ANFIS, proposed by Jang in 1993 [55], is a hybrid model composed of a fuzzy and artificial neural network, where the nodes in the different layers of a network handle fuzzy parameters. This is equivalent to fuzzy inference systems (FIS) with distributed parameters, as shown in figure 7 [55] [52].

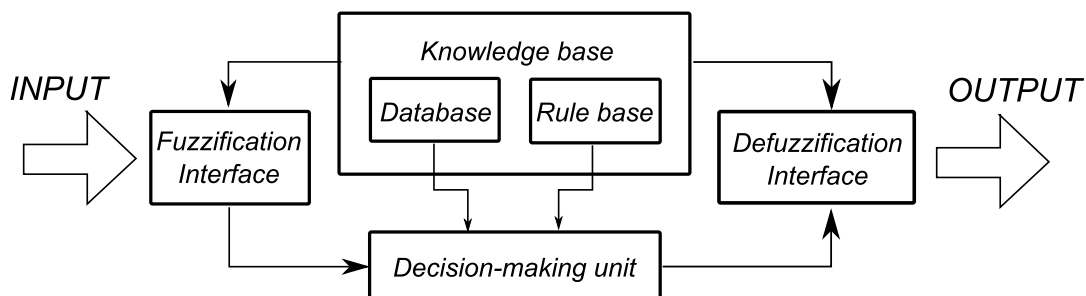


Figure 7. Fuzzy inference system.

Basically, a fuzzy inference system is composed of five functional blocks: 1) A rule base containing a number of fuzzy *if-then* rules. 2) A database which defines the membership functions (MF) of the fuzzy sets used in the rules. 3) A decision-making unit which performs the inference. 4) A fuzzification interface which transforms the crisp (i.e. fixed) inputs. 5) A defuzzification interface which transforms the fuzzy output to the same range of the input.

Usually, the rule base and the database are jointly referred to as the *knowledge base*. For the fuzzification process “membership functions” are used which categorize in a graphic way a specific fuzzy set; it means that a membership function defines a continuous trajectory to where each one of the discrete elements with some property in common belongs.

The steps of fuzzy reasoning (inference operations upon fuzzy *if-then* rules) performed by fuzzy inference systems are:

1. Compare the input variables with the membership functions on the premise part to obtain the membership values (or compatibility measures) of each linguistic label (this step is often called fuzzification).
2. Combine (through a specific T-norm operator, usually multiplication or min.) the membership values on the premise part to get firing strength (weight) of each rule.
3. Generate the qualified consequent (either fuzzy or crisp) of each rule depending on the firing strength.
4. Aggregate the qualified consequents to produce a crisp output (this step is called defuzzification).

Depending on the types of fuzzy reasoning and fuzzy if-then rules employed, most fuzzy inference systems can be classified into three types:

*Type 1:* The overall output is the weighted average of each rule’s crisp output induced by the rule’s firing strength (the product or minimum of the degrees of match with the premise part) and output membership functions. The output membership functions used in this scheme must be monotonic functions.

*Type 2:* The overall fuzzy output is derived by applying “ma” operation to the qualified fuzzy outputs (each of which is equal to the minimum of firing strength and the output membership function of each rule). Various schemes have been proposed to choose the final crisp output based on the overall fuzzy output; some of them are centroid of area, bisector of area, mean of maxima, maximum criterion, etc.

*Type 3:* Takagi and Sugeno’s fuzzy *if-then* rules are used. The output of each rule is a linear combination of input variables plus a constant term, and the final output is the weighted average of each rule’s output.

Figure 8 utilizes a two-rule two-input fuzzy inference system to show different types of fuzzy rules and fuzzy reasoning mentioned above [55].

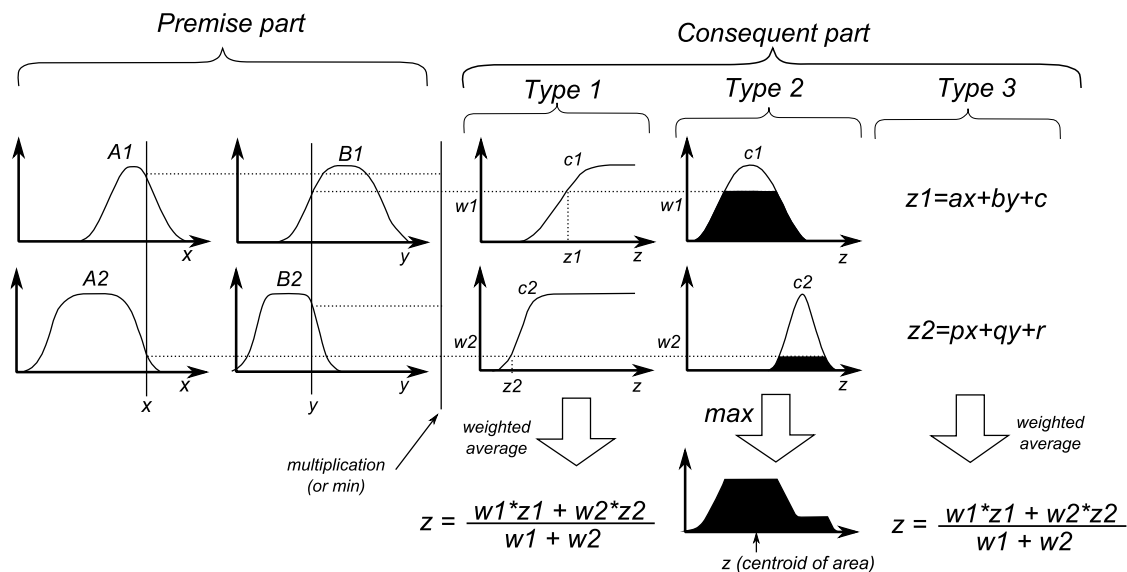


Figure 8. Commonly fuzzy if-then rules and fuzzy reasoning mechanisms.

Thus, ANFIS falls into the third type of fuzzy inference system, where at its core, the technique splits the representation of prior knowledge into subsets in order to reduce the search space and uses the back-propagation algorithm to adjust the fuzzy parameters. The resulting system is an adaptive neural network functionally equivalent to a first-order Takagi-Sugeno inference system, where the input-output relationship is linear.

### 3.2.1 Topology:

An adaptive network is a multilayer feedforward network in which each node performs a particular function (node function) on incoming signals as well as a set of parameters pertaining to this node. The formulas for the node functions may vary from node to node, and the choice of each node function depends on the overall input-output function that the adaptive network requires. Note that the links in an adaptive network just indicate the flow direction of signals between nodes; no weights are associated with the links.

To reflect different adaptive capabilities, both circle and square nodes are used in the representation of an adaptive network. A square node (adaptive node) has parameters while a

circle node (fixed node) has none. The set of parameters in an adaptive network is the union of all parameters of each adaptive node. In order to achieve a desired input-output mapping (detailed later), these parameters are updated according to given training data and a gradient-based learning procedure, although several hybrid techniques can be applied to increase its performance and learning speed.

In a first-order Sugeno system, a typical rule set with two fuzzy *if-then* rules can be expressed as:

$$\text{Rule 1: If } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \text{ then } f_1 = p_1x + q_1y + r_1$$

$$\text{Rule 2: If } x \text{ is } A_2 \text{ and } y \text{ is } B_2, \text{ then } f_2 = p_2x + q_2y + r_2$$

where  $x$  and  $y$  are the crisp inputs to node  $i$ ,  $A_i$  and  $B_i$  are the fuzzy sets in the antecedent,  $f_i$  is the output within the fuzzy region specified by the fuzzy rule; and  $p_i$ ,  $q_i$  and  $r_i$  are the design parameters that are determined during the training process.

In general, ANFIS structure consists of five layers, namely: fuzzy layer, product layer, normalized layer, de-fuzzy layer and total output layer, and a structure for two inputs, one output and the two rules mentioned above is shown in figure 9.

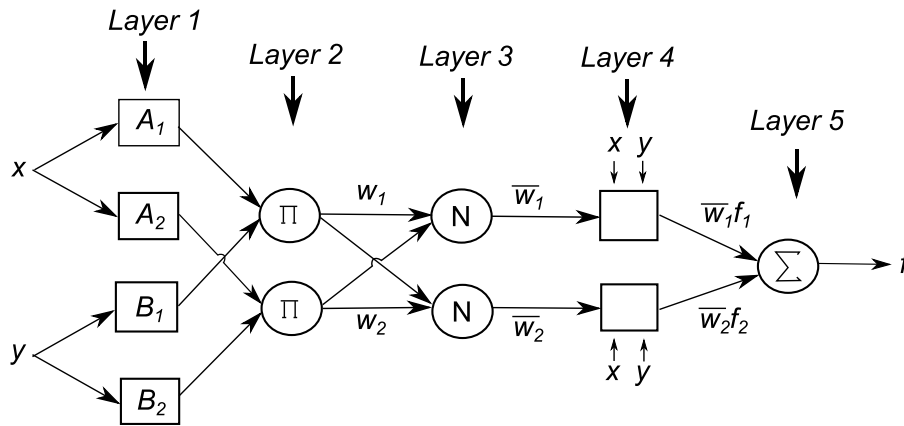


Figure 9. Basic structure of ANFIS.

The characteristics of each layer are explained below.

Layer 1: Every node  $i$  in this layer is a square node with a node function:

$$O_i^1 = \mu_{A_i}(x)$$

Where  $x$  is the input to node  $i$ , and  $A$ , is the linguistic label (small, large, etc.) associated with this node function. In other words,  $O_i^1$  is the membership function of  $A$ , and it specifies the degree to which the input  $x$  satisfies the quantifier  $A_i$ . Usually  $\mu_{A_i}(x)$  is chosen to be bell-shaped with maximum equal to 1 and minimum equal to 0, although it can also be a triangular or trapezoidal membership function, as it is related in figure 10 and whose mathematical definitions are displayed in table 3. Parameters in this layer are referred to as *premise parameters*.

Layer 2: Every node in this layer is a circle node labeled as  $\Pi$  which multiplies the incoming signals and sends the product out. For instance,

$$\omega_i = \mu_{A_i}(x) \times \mu_{B_i}(y), \quad i = 1, 2.$$

Each node output represents the firing strength of a rule (in fact, other *T-norm* operators that perform the generalized function AND can be used as the node function in this layer).

Layer 3: Every node in this layer is a circle node labeled as  $N$ . The  $i$ th node calculates the ratio of the  $i$ th rule's firing strength to the sum of all rules' firing strengths:

$$\bar{\omega}_i = \frac{\omega_i}{\omega_1 + \omega_2}, \quad i = 1,2.$$

For convenience, outputs of this layer will be called *normalized firing strengths*.

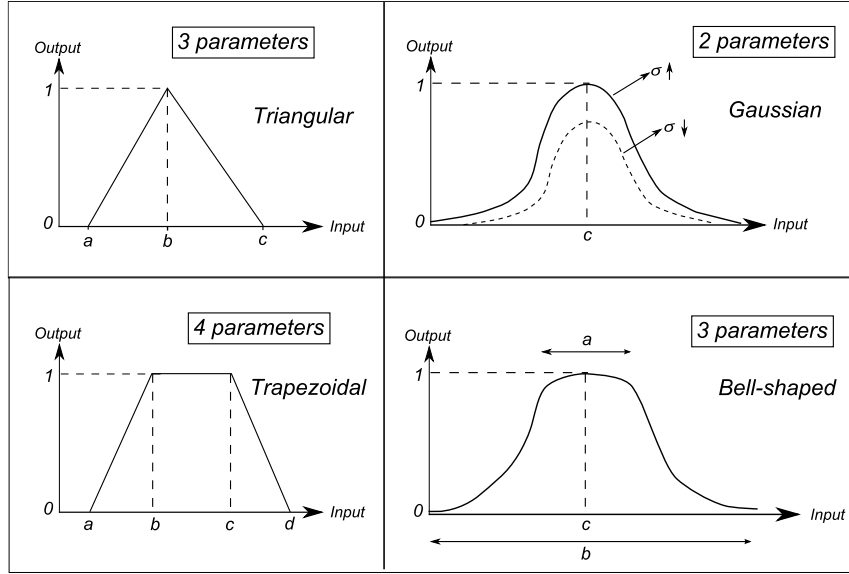


Figure 10. Common membership functions.

Name of MFs	Equation
Triangular MF	$\mu_{Ai}(x) = \max \left\{ \min \left( \frac{x-a}{b-a}, \frac{c-x}{c-b} \right), 0 \right\}$
Trapezoidal MF	$\mu_{Ai}(x) = \max \left\{ \min \left( \frac{x-a}{b-a}, 1, \frac{d-x}{d-c} \right), 0 \right\}$
Gaussian MF	$\mu_{Ai}(x) = e^{-\frac{(x-c)^2}{2\sigma^2}}$
Bell-Shaped MF	$\mu_{Ai}(x) = \frac{1}{1 + \left  \frac{x-c}{a} \right ^{2b}}$

{a, b, c, d} is the parameter set that changes the shapes of the MFs with maximum 1 and minimum 0.

Table 3. Common membership functions.

**Layer 4:** Every node *i* in this layer is a square node with a node function:

$$O_i^4 = \bar{\omega}_i f_i = \bar{\omega}_i (p_i x + q_i y + r_i)$$

where  $\bar{\omega}_i$  is the output of layer 3, and  $\{p_i, q_i, r_i\}$  are the coefficients of a linear combination in the Sugeno inference system. Parameters in this layer will be referred to as *consequent parameters*.

**Layer 5:** The single node in this layer is a circle node labeled as  $\Sigma$  that computes the overall output as the summation of all incoming signals, i.e.:

$$O_i^5 = \text{overall output} = \sum_i \bar{\omega}_i f_i = \frac{\sum \omega_i f_i}{\sum \omega_i}$$

Therefore, an adaptive network which is functionally equivalent to a type-3 fuzzy inference system is constructed, and its fuzzy reasoning is presented in figure 11.

The membership functions distribution on the physical domain for each linguistic variable of the system is called *mapping*, and figure 12 shows that for a 2-input, type-3 ANFIS with nine rules.



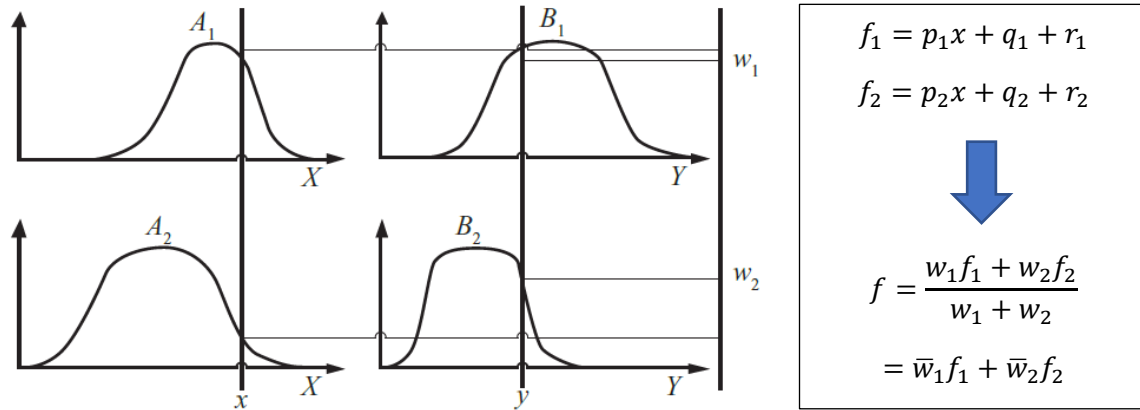


Figure 11. Type 3 fuzzy reasoning (evaluation of the fuzzy inference).

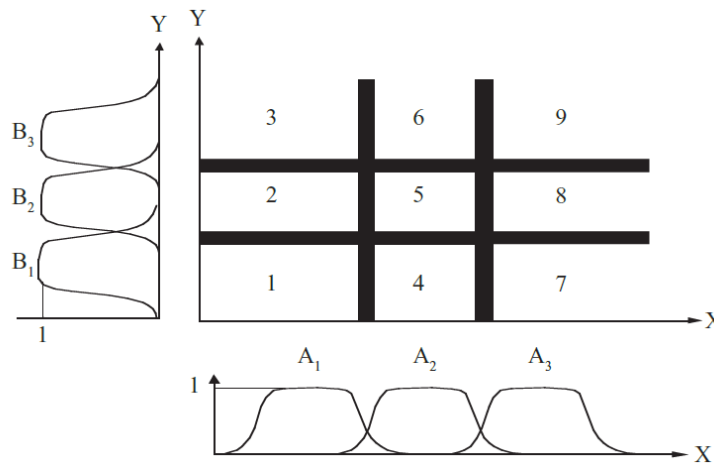


Figure 12. Mapping of a 2-input, type-3 ANFIS with three MF in each input.

Three membership functions are associated with each input, so the input space is partitioned into nine fuzzy subspaces, each one of which is governed by a fuzzy if-then rule. The premise part of a rule delineates a fuzzy subspace, while the consequent part specifies the output within this fuzzy subspace. Considering the above, the number of rules (NR) is given by:

$$NR = \prod_{i=1}^n MF_i$$

Where  $n$  represents the number of inputs of the system.

ANFIS uses a hybrid learning algorithm for the estimation of the premise and consequent parameters. The hybrid learning algorithm estimates the consequent parameters in a forward pass and the premise parameters in a backward pass. In the forward phase, the information propagates forward until layer 4, where the consequent parameters are optimized by a least square regression algorithm. In the backward phase, the error signals propagate backwards and the premise parameters are updated by a gradient descent (GD) algorithm. This measured error is usually defined by the sum of the squared difference between measured and modeled values and is minimized to a desired value. Then, the final overall output in figure 9 can be rewritten as:

$$f_{out} = (\bar{\omega}_1 \cdot x)p_1 + (\bar{\omega}_1 \cdot y)q_1y + (\bar{\omega}_1)r_1 + (\bar{\omega}_2 \cdot x)p_2 + (\bar{\omega}_2 \cdot y)q_2y + (\bar{\omega}_2)r_2$$

Where  $x$  and  $y$  are the input parameters of the model;  $\omega_1, \omega_2$  are the normalized firing strengths of fuzzy rules and  $(p_1, q_1, r_1, p_2, q_2, r_2)$  are the consequent parameters. A more extensive explanation about the learning strategies for ANFIS can be consulted in [55].

# Chapter 4

## Methodology

This chapter is intended to explain the step-by-step of the design and implementation of the indirect estimation model. For this, Chapter 4 is divided in two sections: a first part includes a review about availability of the irradiation data in the Department of Norte de Santander to contextualize the motivation of the present work. In addition to this, an analysis of available climatological information is carried out to identify the parameters whose correlation with the solar radiation can be used in the estimation model. After this, a second section describes the methodology that was used in the development of the ANN and ANFIS models as an alternative information source of irradiation data in the region.

### 4.1. Characteristics of the case study:

Norte de Santander is one of the 32 departments that together with the Capital District of Bogota, form the Republic of Colombia. Cúcuta is the capital city of Norte de Santander and is located in the northeast of the country, in the Andean region, bordering the north and east with Venezuela. The main geographic characteristics are portrayed in figure 13.

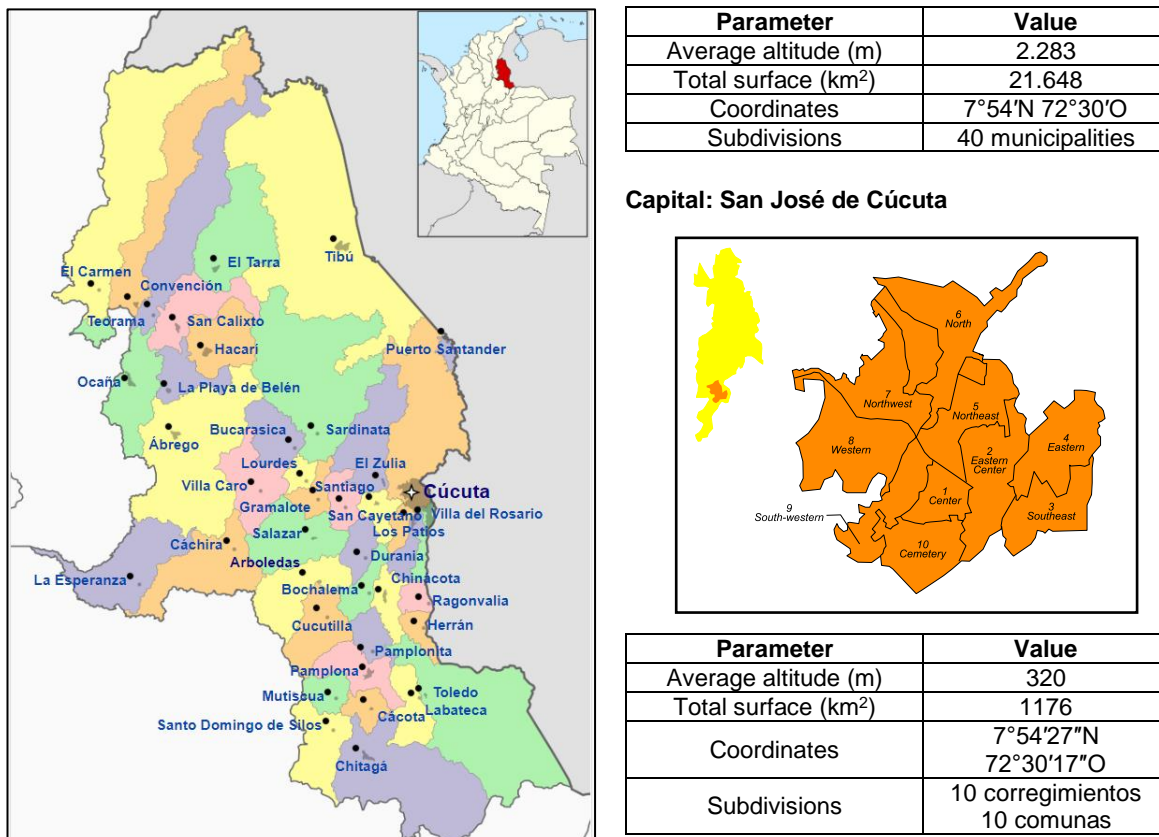


Figure 13. Geographical characteristics of Norte de Santander.

This Department as the rest in Colombia or in a tropical country does not have seasons and therefore, the weather mainly depends on the geographic conditions of the place. The change of altitude between a location and another is one of the principal factors that influences in the climatological variations of the region. Norte de Santander is among the last places crossed by

the Andes Mountain and in consequence, large changes in altitude and thus, in climate, are found between near cities.

For instance, figure 14 and table 4 present the distance and the changes of some meteorological variables among the capital of Norte de Santander and three municipalities (denomination of the administrative territorial division in which the departments in Colombia are organized) in several possible directions. The lines in figure 16 are only a representation of the separation between municipalities but not indicate the real shape of the route.

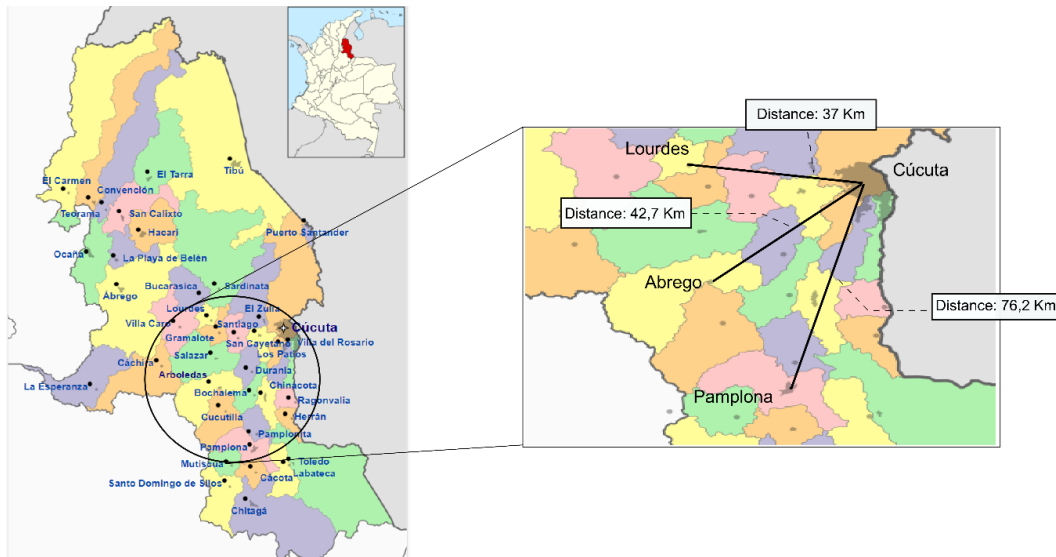


Figure 14. Distances from Cúcuta to different municipalities.

In reference to the information shown in figure 14 and table 4, it can appreciate that a short route of 76 km from Cúcuta to Pamplona (which lasts about 1h 40min in a car according to Google Maps [56]) there is a change in altitude from 320 m to 2.342 m with variations of up to 17 °C in temperature, 6 km/h in wind speed and 34 % for the percentage of humidity in the environment. Similar changes occur for the other two municipalities (see table 4). These significant changes in the altitude due to the mountainous geography of the region in short trips and their effects over the weather conditions, are an indication that the irradiance could also have a particular behavior.

To obtain a wider perspective of the described situation, table 5 shows several climatic variables of the 40 municipalities in the Department of Norte de Santander as for August 27, 2018. The four municipalities used as example in table 4 are highlighted in table 5 for reference.

Considering the high variability in climate conditions, the information sources that provide irradiation data in Norte de Santander must have high accuracy with a spatial distribution small enough to guarantee proper analysis and sizing of renewable energy systems in every municipality, since irradiation values in the region could vary a lot within a few kilometers. This scenario will be covered in more detail in the next sections.

#### 4.1.1. Climatological information sources in Norte de Santander:

For Norte de Santander, several irradiation information sources can be found from either satellite information or empirical models. For the first approach, different databases can be consulted such as: Global Solar Atlas (GSA), NASA Surface Meteorology and Solar Energy (SMSE), and Photovoltaic Geographical Information System (PVGIS). Additionally, empirical models can be analyzed from sources as the PERS project and the IDEAM thorough its website. In this way, in order to identify the benefits of the results of the present research study, these information sources will be used in next chapters for specific cases of sizing and therefore, their main characteristics are displayed below.

Municipality	Delta in altitude (m)	Delta of temperature (°C)	Delta of wind speed (km/h)	Delta of humidity (%)
Lourdes	1.090	9	12	30
Ábrego	1.078	9	12	26
Pamplona	2.022	17	6	34

Note: The differences presented in table 4 were obtained from the average values of each parameter.

Table 4. Climatological variations among Cúcuta and some selected municipalities.

Municipality	Altitude (m)	Minimum temperature (°C)	Maximum temperature (°C)	Delta of temperature (°C)	Minimum wind speed (km/h)	Maximum wind speed (km/h)	Delta of wind speed (km/h)	Minimum humidity (%)	Maximum humidity (%)	Delta of humidity (%)
Arboledas	946	16	23	7	3	8	5	60	93	33
Cucutilla	1277	20	26	6	3	6	3	62	84	22
Gramalote	148	20	26	6	2	9	7	64	75	11
Lourdes	1411	19	26	7	3	9	6	65	79	14
Salazar de Las Palmas	845	17	23	6	5	12	7	53	76	23
Santiago	450	20	27	7	7	28	21	44	69	25
Villa Caro	1600	19	23	4	4	13	9	61	87	26
Cúcuta	320	28	34	6	10	21	11	35	56	21
El Zulia	220	29	35	6	10	19	9	36	56	20
Los Patios	410	27	32	5	10	22	12	40	61	21
Puerto Santander	60	27	34	7	2	12	10	45	74	29
San Cayetano	235	29	35	6	9	18	9	38	57	19
Villa del Rosario	440	27	33	6	10	22	12	41	62	21
Bucarasica	1552	21	27	6	5	14	9	49	72	23
El Tarra	160	27	34	7	2	13	11	44	76	22
Sardinata	320	28	34	6	4	13	9	39	69	30
Tibú	75	27	34	7	3	11	9	44	77	33

Ábrego	1398	20	25	5	3	9	6	59	76	17
Cáchira	2025	23	29	6	5	14	9	50	71	21
Convención	1076	21	27	6	3	11	8	48	83	35
El Carmen	761	20	28	8	4	10	6	49	84	35
Hacarí	1050	19	27	8	2	14	12	50	80	30
La Esperanza	1566	26	32	6	3	9	6	55	76	11
La Playa de Belén	1450	21	28	7	3	14	11	52	76	24
Ocaña	1202	21	27	6	4	9	5	56	78	22
San Calixto	1677	18	24	6	2	12	10	50	82	32
Teorama	72	20	30	10	2	12	10	48	95	47
Cácota	2465	9	16	7	3	12	9	69	93	24
Chitagá	2379	11	17	6	2	13	11	75	95	20
Mutiscua	2600	8	16	8	5	14	9	65	94	29
Pamplona	2342	11	18	7	6	15	9	61	90	29
Pamplonita	1886	16	24	8	6	13	7	56	84	28
Santo Domingo de Silos	2845	6	14	8	4	13	9	70	95	25
Bochalema	1051	19	28	9	3	12	9	52	78	26
Chinácota	1175	18	27	9	4	13	9	54	80	26
Durania	940	19	28	9	3	12	9	48	76	28
Herrán	1995	14	22	8	6	14	8	59	91	32
Labateca	1465	16	24	8	3	14	9	64	90	26
Ragonvalia	1615	14	21	7	5	15	10	56	88	32
Toledo	1642	9	12	3	1	13	12	77	96	19

Table 5. Weather factors for each municipality of Norte de Santander for August 27, 2018.

The available irradiation data from these information sources is described in table 6; the details about the methods and resources that they use for their calculations can be found in: [57] for GSA, [58] for POWER project from SMSE, [59] for PVGIS, [60] for PERS project and in [61] for the IDEAM. All information applied in the next section from these sources was obtained in terms of irradiation ( $\text{Wh/m}^2$ ).

Information source	Irradiation data				Period of availability of the data
	Annual	Monthly	Daily	Hourly	
GSA					2017*
SMSE					1981-2018
PVGIS					2005-2015
PERS					2015
IDEAM					2014*

\*Sources with irradiation data averaged until the date indicated.

Table 6. Irradiation data by information sources and scales.

According to table 6, PVGIS is the most complete information source in terms of scales, with data updated until 2015 (to the date of writing this document). In the specific case of PERS, the empirical models were developed with monthly data, and it is not specified if the results could be adapted to other scale. For this reason, we just considered the data supplied by the authors in the official website of the PERS project (i.e., information for 2015 in monthly scale). It is important to indicate that the information supplied by this source is in a graphical way through a Cartesian plane, which relates the interception points of the irradiation values and its corresponding evaluation period (month); therefore, this scenario can generate inaccuracies in the acquisition of the data used in the subsequent analysis, since the numerical data are not available.

Regarding GSA and SMSE databases, the irradiation data is available up to 2017 but not in all scales. Besides, GSA has an additional disadvantage since does makes available just annual data averaged from its first acquisition until 2017, i.e., does not present information year after year in order to analyze the changes from a period to another.

Finally, the IDEAM published a solar atlas which contains a set of maps showing the annual and monthly irradiation averages. To generate those maps, IDEAM used several empirical models that use irradiation and sunshine data from the meteorological network of Colombia; one disadvantage of this source is that both the annual and monthly data are given in irradiation ranges (for instance, the annual map indicates that the irradiation for Cúcuta is between  $4 \text{ kWh/m}^2$  and  $4,5 \text{ kWh/m}^2$ , i.e., there is a range of uncertainty of  $0,5 \text{ kWh/m}^2$ ) so that the accuracy of the information is low. In addition to this, as the GSA data, the IDEAM does not present historical data but an average from the last years recorded by the entity until 2014. Despite of this, information more detailed can be obtained from the IDEAM but only for some cities of the country; this information is based on irradiation data in hourly scale measured in several climatological stations (as it will be shown later), but that just covers a small part of the national territory. This data is not available in the website as the atlas, since it must be requested after going through a series of administrative processes with the entity.

#### 4.1.2. Climatological information available in Norte de Santander:

As it has been explained previously, different climatological variables or sky images can be used for forecasting solar radiation in a specific place; since the current case study is the Department of Norte de Santander, the input variables for the model will be limited by those measured by the different weather stations in the region.

The entity responsible for managing the weather stations is the IDEAM which has global solar radiation information from a network of Conventional Stations (with measurements mainly from actinographs and a few pyranometers) and other Automatic Satellite Stations (with data from

pyranometers). Pyranometers installed from 2005 by the IDEAM, are the devices that are currently measuring irradiance in the country, since it was decided to dismantle the actinographs in operation, due to the difficulties encountered in the evaluation of their results (graphs based on a special cardboard) [62].

In total the IDEAM has 4.504 stations which can be classified in conventional and automatic satellite [63], from the total only 83 measure irradiation data [64]. Nowadays, the IDEAM also performs the measurement of the following variables continuously: temperature (minimum, maximum and average value), humidity (relative humidity, vapor tension and evaporation), wind (speed and direction), precipitation, insolation (sunshine) and amount of ozone. According to the number of variables recorded by the stations, and their sample frequencies, these are classified in several categories related in table 7 [65].

Category	Description
Main synoptic	It allows to observe hourly meteorological variables such as cloudiness, direction and wind speed, atmospheric pressure, air temperature, type and height of the clouds, visibility, special phenomena, humidity, precipitation and extreme temperatures.
Main climatological	Observations of visibility, present atmospheric time, quantity, type and height of the clouds, soil condition, precipitation, air temperature, humidity, wind, solar radiation, solar brightness, evaporation and special phenomena are measured in this station.
Ordinary climatological	This type of stations necessarily has a rain gauge and a psychrometer. Thus, they can measure rainfall and instantaneous temperatures. Other variables can be also present in the measurements of this station.
Agrometeorological	In this category, meteorological and biological observations are made, including phenological data and other observations that help to determine the relationships between weather and climate, on the one hand, and the life of plants and animals, on the other. It includes the same program of observations of the main climatological station, plus temperature records at various depths (up to one meter) and in the layer near the ground (0, 10 and 20 cm above the ground).
Pluviometric	It is a meteorological station equipped with a rain gauge or container that allows to measure the amount of rainfall between two consecutive observations.
Pluviographic	It records the precipitation, in a graph way that allows to know the amount, duration, intensity and period in which the rain has occurred. Currently, daily recorders are used.

Table 7. Main categories of the meteorological stations of the IDEAM.

The specific meteorological scenario for Norte de Santander is as follows: the IDEAM has 220 stations located in the Department, of which 39 are classified among the first four categories in table 7, where several measurements of interest for a solar estimation model are recorded (temperature, wind speed, humidity, etc.). Therefore, the other stations located in the rest of categories in table 7, are not relevant for the current analysis because they only measure one variable, as the pluviometric stations. This is, as mentioned in chapter 2, because the use of several weather parameters improves the accuracy of the model based on ANN.

In addition to the above, of the 39 aforementioned stations, 14 stations are suspended by different maintenance programs or technical problems; thus, 25 stations can be used to obtain meteorological data in Norte de Santander. These stations are listed in figure 15a and detailed in table 8.

Name	Category	Municipality	Location	Altitude (m)
Apto. Camilo Daza	Main synoptic	Cúcuta	(7,93, -72,50)	250
ISER Pamplona	Agrometeorological	Pamplona	(7,37, -72,64)	2,34
La Esperanza	Ordinary climatological	Ragonvalia	(7,56, -72,53)	1,76
P. Nacional El Tama	Main climatological	Herrán	(7,42, -72,44)	2,5
Ragonvalia	Main climatological	Ragonvalia	(7,57, -72,48)	1,55
Univ. Francisco de Paula Santander (UFPS)	Main climatological	Cúcuta	(7,89, -72,48)	311
Univ. de Pamplona	Agrometeorological	Pamplona	(7,36, -72,66)	2,362
Alcaldía De Herrán	Main climatological	Herrán	(7,50, -72,48)	2,04
Finca La Palmita	Ordinary climatological	Pamplonita	(7,51, -72,64)	1,107
Apto. Camilo Daza (Automatic)	Main synoptic	Cúcuta	(7,93, -72,51)	313
Carmen de Tonchala	Main climatological	Cúcuta	(7,84, -72,56)	285
Salazar	Main climatological	Salazar	(7,77, -72,83)	860
Cinera-Villa Olga	Main climatological	Cúcuta	(8,16, -72,46)	100
Tibú	Ordinary climatological	Tibú	(8,63, -72,72)	50
Sardinata	Ordinary climatological	Sardinata	(8,07, -72,80)	320
Apto. Aguas Claras	Main climatological	Ocaña	(8,31, -73,35)	1,435
Teorama	Ordinary climatological	Teorama	(8,44, -73,28)	1,16
Abrego Centro Adm.	Main climatological	Ábrego	(8,08, -73,22)	1,43
La Playa	Ordinary climatological	La Playa	(8,21, -73,23)	1,5
Ins Agr Convencion	Main climatological	Convención	(8,47, -73,34)	1,076
UFPS. (Automatic)	Ordinary climatological	Ocaña	(8,23, -73,32)	1,15
Aguas De La Virgen	Main climatological	Ocaña	(8,22, -73,39)	1,7
Esc Agr Cachira	Ordinary climatological	Cáchira	(7,73, -73,05)	1,882
Silos	Ordinary climatological	Silos	(7,20, -72,75)	2,765
Tunebia	Ordinary climatological	Toledo	(7,00, -72,11)	370

Table 8. Active meteorological stations for Norte de Santander.

From table 8, a total of 7 stations (4 conventional and 3 automatic satellite ones) measure irradiation but since the conventional stations with actinographs are becoming dismantled, just the 3 automatic satellite stations are enabled to provide irradiance data in a reliable way. These 3 stations are shown in figure 15b and highlighted in table 8.

Thus, the stations with measurements of irradiation only constitute 1,36 % of the total stations installed in Norte de Santander. This confirms for this region of Colombia, the common scenario in several countries around the world mentioned in the first chapter of the present document, in reference to the low ratio between weather stations with measurements of climatological variables as humidity, temperature, etcetera, and those with irradiation data.

Other weather stations with ground measurements (in total 11) can be found in the Department, managed by the Regional Autonomous Corporation of the Northeast Frontier (CORPONOR, by its acronym in Spanish). However, in [60] was shown that the databases of nine (9) of the eleven (11) stations did not present radiation information or insolation. In addition, the two (2) remaining ones did not contain information for a period of time greater than five (5) years with absence of information in several months, for which it was determined that this information is not representative for the analysis of the solar potential of the Department.



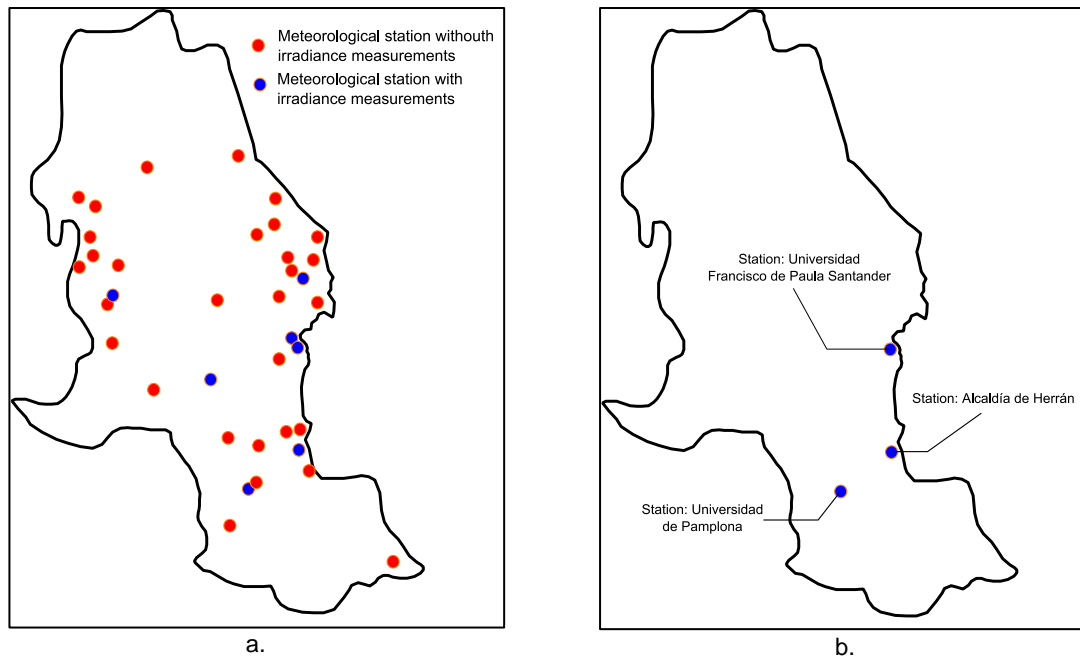


Figure 15. Meteorological stations for Norte de Santander: a. Classified in the first four categories from table 7. b. Only those with reliable irradiation data.

Considering the previous sections, most of the territory of Norte de Santander does not have solar information directly measured so PV sizing must be supported by data estimated from different sources. Here is where the accuracy of the estimation starts to play a significant role based on the geographical variability of the region and the diversity of its weather conditions. Thus, taking into account the good results of the estimation models obtained with ground-measurements over the satellite ones (for the specific case of Norte de Santander, this choice can be validated by the work developed in [45] which presents some evaluation criteria to the selection between a ground-based or satellite measurements in countries with a landform and weather conditions very similar to Colombia), and the profits achieved by the AI techniques among the ground-based methods, it makes sense to use indirect estimation models with AI in Norte de Santander. This is the focus of the rest of the document, where the construction of the AI models for the solar estimation is performed, and later, the results are compared under several scenarios in order to identify the relevance of their implementation.

#### 4.2. *Development of the model:*

The main idea of the present work is to design an estimation model for irradiation data which takes advantage of the correlation between several climatological variables (as temperature humidity, sunshine, among others) of easy access and the solar radiation, to increase the information available in Norte de Santander and consequently to foster the implementation of PV systems in the region. Basically, the objective of the model is to define a mathematical relationship from specific input data (climatological variables) and output values (irradiation data) with high accuracy. For this reason, for the construction of the input-output links of the model, the AI techniques selected require a training process with the input data and its expected output.

In this sense, as only three municipalities have climatological stations with irradiation measurements for the training and validation process of the networks, just three models can be constructed for their analysis.

Considering the above, three models will be generated and evaluated in terms of the estimation accuracy to guarantee the reliability of their results. Later, these will be implemented in other zones with similar weather and geographical conditions to define their degree of adaptability in places where irradiation measurements are not available. For this reason, the present study seeks

to be a contribution to the results of the PERS project, considering that a similar methodology is applied, but using AI models instead of empirical ones, because different authors have reported in the literature that AI models have superior performance than empirical models.

Thus, as the locations without irradiation data will use the indirect estimation models to predict the solar information, these must employ input variables present in the largest number of stations in order to cover a wider zone of the Department. Therefore, the first filter to identify the climatological variables that will be applied as input in the construction of the model is their availability among the 25 weather stations indicated in the previous section; the second filter is selecting the variables that have demonstrated to have high correlation with solar radiation and hence, their use is extensive in the literature.

After a thorough revision of the stations and research in the literature according to the previous filters, the variables selected to participate in the construction of the model were: humidity, temperature, sunshine, and wind speed.

Having defined the input variables of the model, the main steps for its construction are the following:

1. Pre-processing of the data.
2. Selection of the network topology, training and validation of the models.

The previous steps are detailed in the next sub-sections.

#### 4.2.1. Pre-processing stage:

In some occasions, the information recorded for the different climatological variables presents errors such as missing data, over-peak or atypical measurements mainly related to malfunction of the measuring instruments [53]. This can carry out inaccuracies in the construction of the model, so that correction techniques must be applied to delete those errors and guarantee a correct analysis of the data. In addition to this, not only errors must be identified but problems in the structure of the data, consequently, several organization algorithms are developed to solve it. Before explaining the techniques and algorithms used in this stage, the data obtained for this research study are presented.

Data from 2005 to 2017 with intervals of 1 hour were obtained from IDEAM. The number of data points in hourly intervals for each variable are presented in table 9. The period between samples was established based on the minimum interval available from the IDEAM for most parameters under evaluation.

Station	Radiation	Humidity	Wind speed	Temperature
Universidad Francisco de Paula Santander	73.617	81.708	392.470	76.576
Universidad de Pamplona	32.250	35.781	206.093	35.756
Alcaldía de Herrán	38.759	52.465	350.288	61.293

Table 9. Distribution of the amount of data provided by the IDEAM.

The sensor used by the IDEAM for the irradiation measurements is the CM11 from the company ADOLF THIES GMBH & Co. KG. As it can be observed from table 9, the amount of data points for the variable *wind speed* is much larger than the one presented for the other variables; the reason for that is because wind speed was provided in intervals of 10 minutes instead of 1 hour as the other ones.

Based on the above, an average of the data for each hour was calculated in order to obtain the same period for all variables. It is worth to mention that although several data points were not

provided in the file supplied by the IDEAM (in a similar way to the other variables), in this case no technique to replace those values was implemented, assuming that the changes between the intervals of 15 minutes were too small so that their absence to affect the final average. The final amount of wind speed data used for the pre-processing stage were: 68.688, 35.882 and 60.256 for the stations in Cúcuta, Pamplona and Herrán, respectively.

In this sense, sunshine data were not provided directly from IDEAM since some weather stations with irradiation measurements, do not record that variable. For this reason, the data for that variable were calculated from the irradiation data considering the concept of sunshine duration given by the IDEAM in [66], where it defines sunshine as the sum of the subperiods during which the direct solar irradiance exceeds  $120 \text{ W/m}^2$ . These hourly calculations were compared with daily data from other stations in the same city to guarantee the reliability of the process.

After presenting the data to analyze in the pre-processing stage, the different steps with their corresponding algorithms to perform this stage are described below. From this section on, all the algorithms for the manipulation of data were implemented using the software Matlab R2017a from Mathworks which is a well-known and specialized tool for this type of processing.

- *Step 1: Organization of the data*

Data from IDEAM were supplied in a *.txt file* which was converted into spreadsheets in Microsoft Excel Professional 2016 to facilitate the access from several algorithms in Matlab.

In this first step, the set of data were organized taking into account the following considerations:

1. The experimental data or measurements were limited in a time range from 7:00 to 18:00 by day, because solar radiation exists only in this interval. This is due to the absence of seasons which causes that solar radiation has presence in a constant hourly range during the whole year.
2. The experimental data or measurements not recorded in the original file were included with a value of zero, except if the missing data covered the whole day. In the latter case, a notification is generated, considering that days with more of two missing data are deleted in the next steps (which is detailed later).
3. The experimental data or measurements are distributed sequentially, so that any data out of its corresponding set was re-located (either in the course of the day or month).

A flowchart of the algorithm used for this first step is displayed in figure 16.

- *Step 2: Interpolation and extrapolation of missing data*

After distributing the data sequentially in sets of 12 values per day (range from 7:00 to 18:00), this step must analyze which of these sets represent a reliable data sample; for this, a similar concept to the one used in [53] was applied, to determine the maximum amount of missing data that can be tolerated in a specific group. There, it is posed that if close 16,6 % of the data of the set were lost, this set of data should be deleted. Adapting this criterion to the sets of data from step 1, only 2 missing data can be allowed in each set (irradiance hours in a day), being this limit the 16,6 % of the sample under evaluation. In this way, the considerations in the algorithm applied in this step are the following:

1. Days with more than 2 missing data are deleted of the total group of data, and these are reported in a *excel file* for future consultation.
2. Days with 2 missing data or less use interpolation or extrapolation techniques according to the case to fill them and re-build the irradiance profile of the day [17]. The interpolation method implemented in the algorithm was the *Piecewise Cubic Hermite Interpolating*

*Polynomials (PCHIP)* [53] and for the extrapolation process a *Polynomial Regression* based on data modeling in Matlab [67] was applied. The results are also exported in a *excel file* to continue with their processing in the next step.

A wider description of the interpolation and extrapolation techniques is given in [68], [69] and [70].

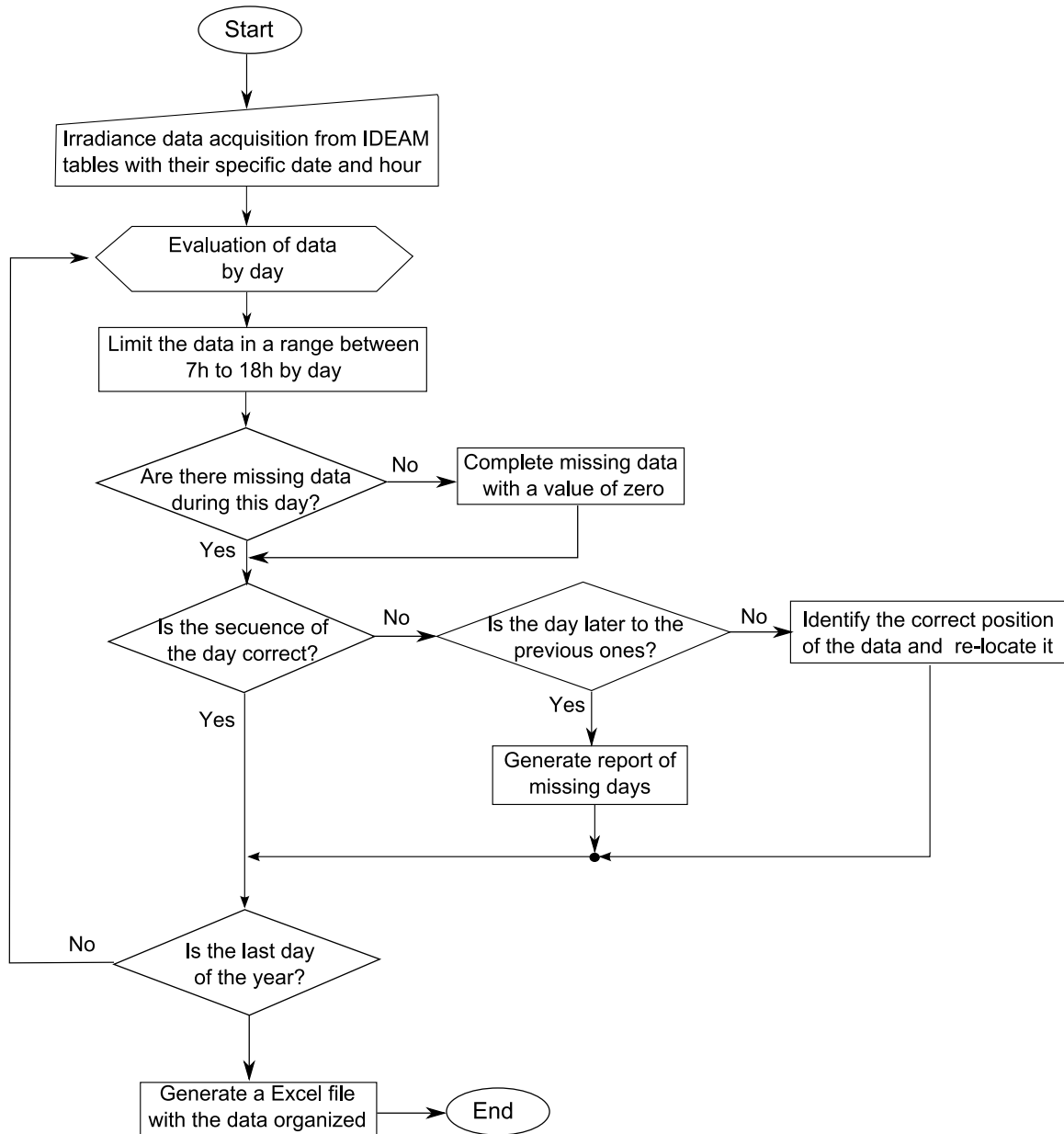


Figure 16. Flowchart of the algorithm in Matlab for the first step of the preprocessing of data: Organization.

The amount of previous or subsequent data used in each process (both in the interpolation and extrapolation) depended on the location of the missing data. It is needed much more data when these were adjacent; and if additionally adjacent missing data were located in the limits of the day, i.e., at 7h or 18h, both the interpolation as the extrapolation process should be applied.

Figure 17 shows the flowchart of the algorithm in Matlab used in this step. Finally, the results of the second step are presented in tables 10, 11, and 12. Table 10 shows the amount of deleted days by the algorithm; table 11 presents the percentage over the total amount of data for the interpolated and extrapolated data points; table 12 contains the final amount of data that the outlier detection algorithms will analyze in the next steps.

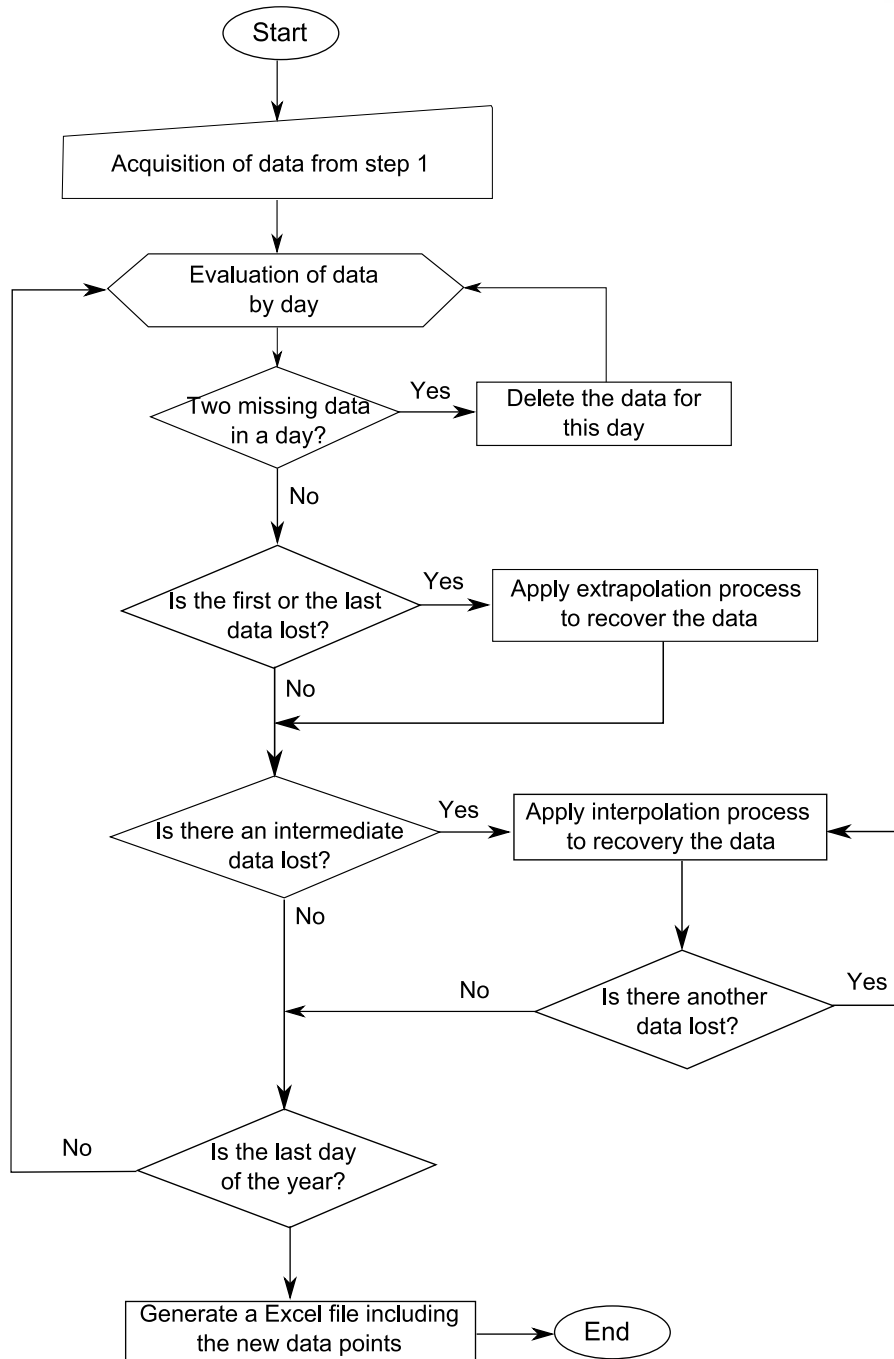


Figure 17. Flowchart of the algorithm in Matlab for the second step of the pre-processing of the data: Interpolation and extrapolation processes.

It is important to indicate that table 12 presents the final amount of data that are evaluated by the methods to detect atypical values, however, the irradiation data must still be analyzed by the extraterrestrial solar radiation criterion (explained later) before being processed by those algorithms.

Station	Irradiation	Humidity	Wind speed	Temperature
Universidad Francisco de Paula Santander	966	515	314	545
Universidad de Pamplona	138	180	131	255
Alcaldía de Herrán	202	495	202	213

Table 10. Days deleted from organized data because of the excess of missing data.

Station	Irradiation	Humidity	Wind speed	Temperature
Universidad Francisco de Paula Santander	4,62 %	1,79 %	0,73 %	1,81 %
Universidad de Pamplona	1,04 %	1,02 %	0,41 %	1,03 %
Alcaldía de Herrán	1,40 %	1,33 %	0,38 %	0,90 %

a.

Station	Irradiation	Humidity	Wind speed	Temperature
Universidad Francisco de Paula Santander	1,00 %	0,50 %	0,39 %	0,92 %
Universidad de Pamplona	0,95 %	0,30 %	0,30 %	0,30 %
Alcaldía de Herrán	1,87 %	2,10 %	0,47 %	0,77 %

b.

Table 11. Percentage of missing values: a. Interpolated. b. Extrapolated.

Station	Irradiation	Humidity	Wind speed	Temperature
Universidad Francisco de Paula Santander	30.312*	37.884	33.492	35.904
Universidad de Pamplona	15.540*	17.004	17.496	16.104
Alcaldía de Herrán	19.368*	27.396	30.012	30.408

\*Values modified in subsequent step.

Table 12. Total amount of data points for the outlier detection algorithms.

### - Step 3: Handling of atypical values

To identify erroneous or outlier data, two methods were applied: analysis of standardized residuals [53] and the Chauvenet criterion. In addition to this, incorrect records of the global solar radiation are revealed using daily clearness index ( $K_t$ ) as an indicator.  $K_t$  is calculated as the ratio of daily global solar radiation intensity measured to the daily extraterrestrial solar radiation on a horizontal surface, being the latter defined as the beam of nearly parallel incident sunrays on top of the Earth's atmosphere, before penetrating it and suffering losses by the different factors present in its trajectory to the ground [71]. The upper and lower limits for  $K_t$  represent a clear sky and completely cloudy sky, respectively. When the daily clearness index was outside the range of 0,015 to 1, the data were considered erroneous and deleted. This analysis is applied to the irradiation data before using the outlier detection. Each one of these methods are explained in the next sections.

#### 1. Analysis of Standardized Residuals:

It is a measure of the intensity between the difference of an observed and expected (estimated) value. It is a normalization method used to compare the residual data of a model; this, since typically the standard deviations of residuals in a sample vary greatly from one data point to another even when the all errors have the same standard deviation (particularly in regression analysis). Thus, it does not make sense to compare residuals at different data points without first standardizing. Standardized residual is a statistical process very common for the detection of the *outliers*, and it can be used as a general rule based on the following considerations [72]:

- If the standardized residual is negative so the observed data is much less than the expected value; this concept of "negative" can be limited for values smaller than  $-2$ .
- If the residual is positive so the observed data is much greater than the expected value; in a similar way to the previous concept, "positive" can be limited for values higher than  $2$ .

These statements which constitute the rule for the standardized residual technique is analog to the empirical rule: 68 95 99,7 for normal distributions, where the 95 % of the data are inside of the area limited to two standard deviations from the average of the data [68]. Then, if the residual is  $\pm 3$ , this means that it is a rare event. Mathematically, the concept is given by equation 1 [73].

$$\text{Standardized residual} = \frac{\text{Residual}}{\text{Standard deviation of residual}} \quad \text{Eq. 1}$$

Where the *residual* is the difference between the observed data (experimental) of the independent variable  $y$  and the expected data estimated by fitting a model as a linear or polynomial regression  $\hat{y}$ :

$$\text{Residual} = y - \hat{y}$$

Note that the term *residual* is different to the one of *error*, since the latter is considered as the difference between the observed data and the real value of an amount of interest (for instance, the population mean), and not by an estimated value (for instance, the sample mean) as in the *residual* concept [74].

In Matlab, the application of the *standardized residual process* is supplied for the detection of the outliers with linear regression assumptions; for this case, assuming that the *Residual* matrix is a table of  $n$  by 4 containing four types of residuals, with a single row for each observation, the *standardized residual* is defined as the set of raw residuals divided by their estimated standard deviation [75]. For a specific observation  $i$ , the *standardized residual* appears as:

$$st_i = \frac{r_i}{\sqrt{MSE(1 - h_{ii})}}$$

Where *MSE* is the mean squared error and  $h_{ii}$  is the leverage value for observation  $i$ , and it is handled by the object *mdl* for the fitting of the model from the functions *fitlm* or *stepwiselm* in collaboration with the dot notation *mdl.Residual.raw*. On the contrary, for polynomial regressions the analysis suggested is based on the *Curve Fitting app* and in its graphic results which can be obtained through the option *Residual Plot*, as it is presented in [76].

Therefore, to apply the concept of *standardized residual* in data whose fitting models (estimated values) for calculating the residuals, need different methods to a *linear regression* (as it is the case of the current work), Matlab requires the development of an own and custom algorithm for their handling. For this reason, an algorithm for the detection of outliers considering the characteristics of this statistical method was developed with the following considerations:

- Data sets of 12 values from step 2 were acquired, and a polynomial regression fitting to the daily irradiance profile was performed.
- The corresponding *residual* for each data from the previous point was determined, and the standard deviation for these residual values was defined.
- The standardized residual for each data based on equation 1 is obtained, and subsequently it is compared with the limits mentioned in the general rules for this concept, i.e., if the result was higher than 2 or smaller than -2, the data was considered as an atypical value and later, it was deleted. A report about this was generated with the date and value of the analyzed variable. The data deleted were interpolated or extrapolated according to its position.

As an example of the above, the irradiance data point at 15:00 on January 2, 2014 in Cúcuta was cataloged as an atypical value by the standardized residual algorithm, and the detection and modification process of this data was as follows: the experimental data for this day, i.e., data given as result of the interpolation and extrapolation algorithm were used to obtain the coefficients of a polynomial structure based on a regression process, which act as the expected values mentioned

in the definition of the concept of standardized residual; both the experimental and expected data can be visualized in figure 18a. The residual for each data is calculated and presented in table 13. The standard deviation for the set of residuals of the table 13 is calculated considering equation 2.

$$\sigma = \sqrt{\frac{\sum_{i=0}^{N-1} (r_i - \hat{r})^2}{N - 1}} \quad Eq. 2$$

Where  $r_i$  represents the residual in the position  $i$ ,  $\hat{r}$  is the average residual and  $N$  is the number of the data point, which for this case is 12. The standard residual for each hour is presented in table 14, where it can be observed that the corresponding irradiation data point at 15:00 on January 2, 2014 (highlighted) exceeds the limits established as criterion in the outlier detection.

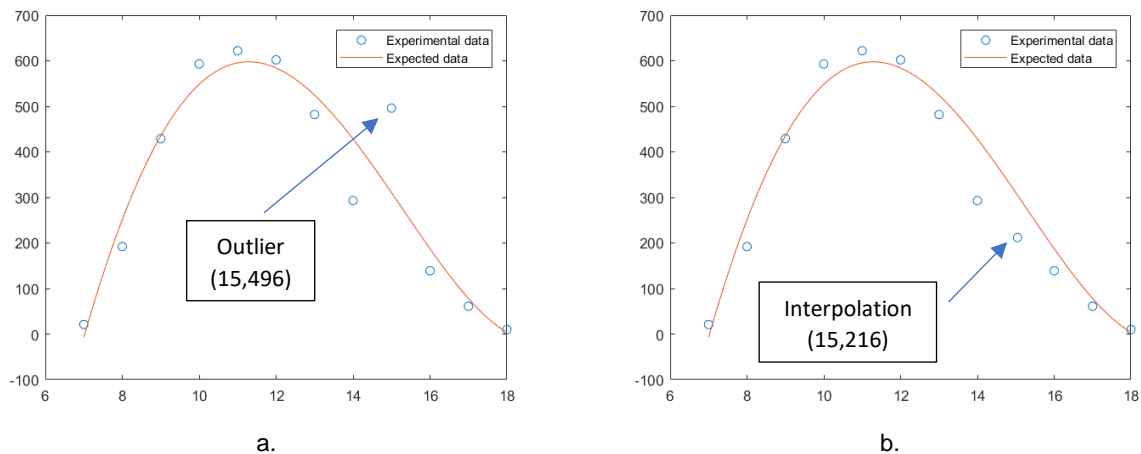


Figure 18. Standardized residual process: a. Detection of outlier. b. Correction of outlier.

Hour*	7	8	9	10	11	12	13	14	15	16	17	18
E. D. <sup>1</sup> (Wh/m <sup>2</sup> )	21	192	429	593	722	602	482	293	496	139	61	10
F. D. <sup>2</sup> (Wh/m <sup>2</sup> )	-7,04	251,2	435,3	548,3	595,4	584,1	524,4	428,3	310,4	187,3	78,1	4,1
Residual (Wh/m <sup>2</sup> )	28,04	-59,2	-6,34	44,7	26,6	17,9	-42,4	-135	185	-48,3	-17	5,9

\* Hour in 24h format.

<sup>1</sup> Indicative for: Experimental data.

<sup>2</sup> Indicative for: Fitting or expected data.

Table 13. Example data for the application of the standardized residual criterion.

Hour*	7	8	9	10	11	12	13	14	15	16	17	18
S. R. <sup>1</sup>	0,36	-0,77	-0,08	0,58	0,34	0,23	-0,55	-1,76	2,41	-0,62	-0,22	0,07

\* Hour in 24h format.

<sup>1</sup> Indicative for: Standardized residual.

Table 14. Standardized residuals for data in table 13.

After outlier detection, this is deleted and subsequently interpolated as it is shown in figure 18b. This process is cyclical for all days of the years evaluated, and for each one of the variables under analysis (humidity, temperature, wind speed, sunshine and irradiation). An overall flowchart of the standardized residual algorithm that was applied is displayed in figure 19.

At this point, it is worth to mention the degree of the polynomial used in the regression process; for its selection, several samples (three for each month of the year by all the years and stations)



were analyzed in the *Curve Fitting Tool* of Matlab in order to identify the degree with the best results for each variable and to be able to implement it in general form for all data under evaluation. At the end of this process, a 4<sup>th</sup> degree polynomial regression was applied because it presented average R-square values close to 1 and average Sum of Squares due to Error (SSE) values close to 0 (as it is shown in table 15), which indicates that the model has a smaller random error component, and that the fit will be more useful for prediction [77].

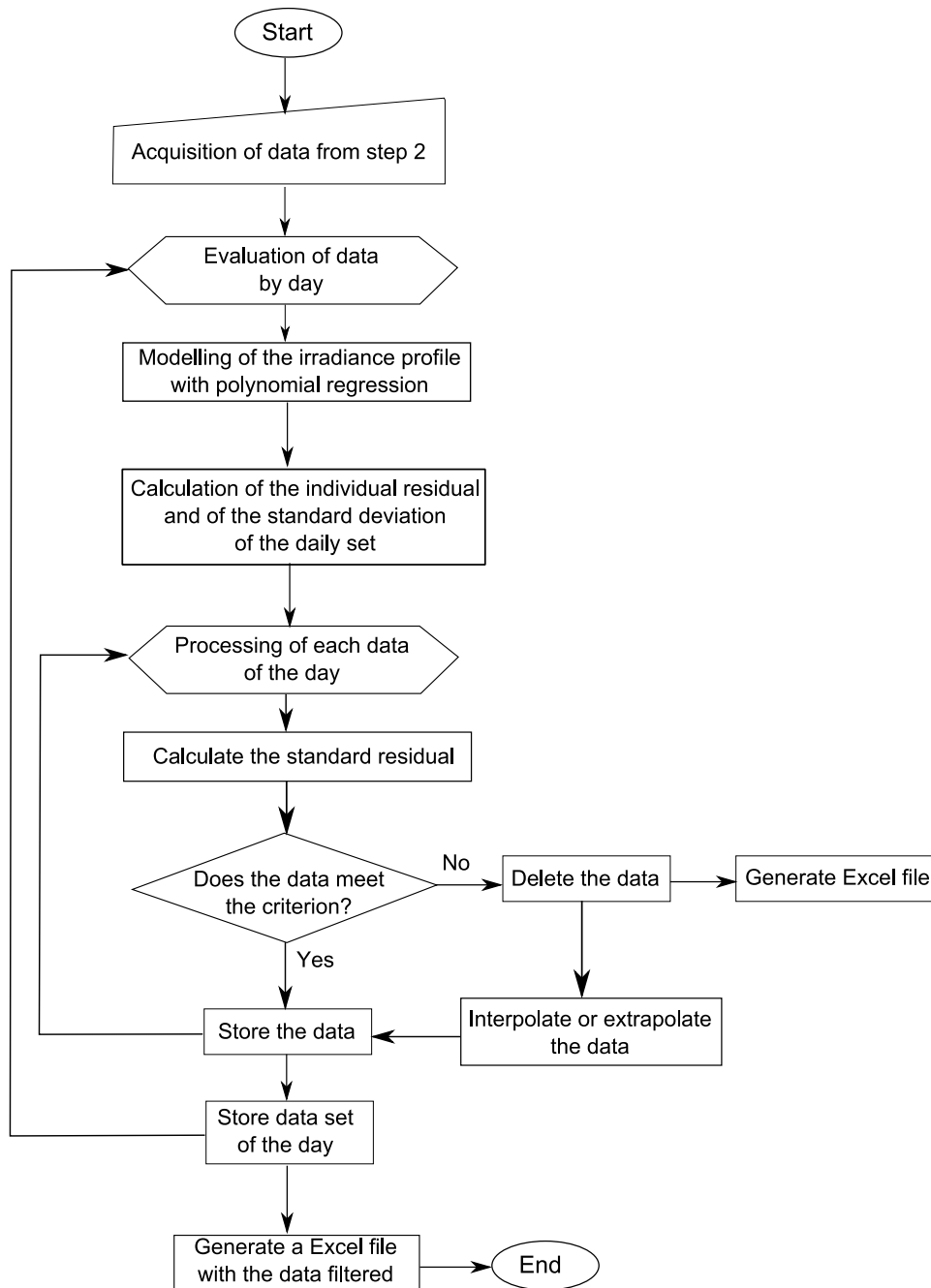


Figure 19. Flowchart of the standardized residual algorithm implemented in Matlab.

Variable	Average R-square	Average SSE
Irradiation	0,9871	17,02
Humidity	0,9732	4,87
Temperature	0,9741	1,79
Wind speed	0,9845	2,03

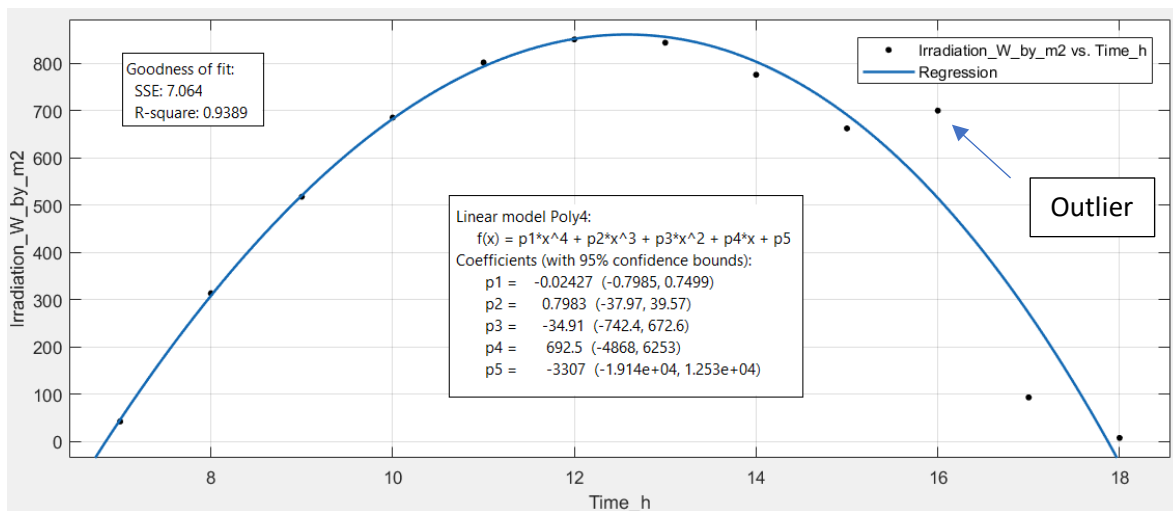
Table 15. Fit indicators for each variable under evaluation.

In a similar way, a degree higher than 4 showed a better R-square and SSE value but the generated regressions were more sensitive to atypical values as it is shown in figure 20 for a set of irradiation data; only with the participation of one atypical value, the model calculated by the regression presented large variations looking for the best fit to the experimental values instead of determining a model for the expected values. This was the other criterion to define a 4<sup>th</sup> degree polynomial regression instead of one with a higher degree. A 4<sup>th</sup> degree regression maintains the equilibrium between the indicators which show a good fit and the stability to atypical values.

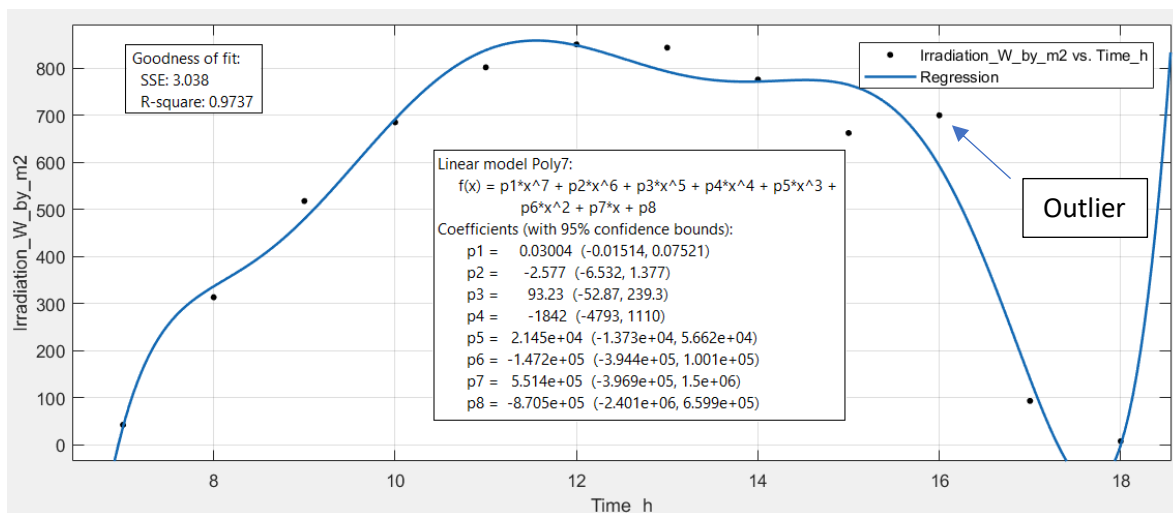
The results of the analysis of standardized residuals are presented in table 16 for each of the stations.

Station	Irradiation	Humidity	Wind speed	Temperature
Universidad Francisco de Paula Santander	5,18 %	5,08 %	4,47 %	5,24 %
Universidad de Pamplona	5,12 %	4,90 %	4,52 %	4,96 %
Alcaldía de Herrán	5,24 %	5,20 %	4,47 %	5,05 %

Table 16. Percentage of modified data by the analysis of standardized residuals.



a.



b.

Figure 20. Polynomial regression for a set of irradiation data: a. With 4th degree. b. With 6th degree.

## 2. Chauvenet criterion:

Chauvenet's criterion is a method that can identify the bad data or wild point and discard them from experimental series. In [78], it was used as atypical values detection criterion for a polynomial parametric equation obtained from experimental data, similar to the application which is looked for in the current work. Other investigations have also used this criterion as an outlier detection mechanism for different types of experimental data [79] [80].

Chauvenet's criterion states that if the expected number of measurements is at least as deviant as the suspect measurement (where the term "suspect" makes reference to the possible atypical value in evaluation) is less than one half, then the suspect measurement should be rejected [81]. For instance, for a set of data of  $N$  measurements:

$$x_1, \dots, x_N$$

of a single quantity  $x$ . From all  $N$  measurements, it can be calculated  $\hat{x}$  (mean of the measurements) and its standard deviation  $\sigma_x$ . If one of the measurements (called  $x_{sus}$ ) differs from  $\hat{x}$  so much that it looks suspicious, then find:

$$t_{sus} = \frac{|x_{sus} - \hat{x}|}{\sigma_x} \quad Eq. 3$$

the number of standard deviations by which  $x_{sus}$  differs from  $\hat{x}$ . Next, from Appendix A, we can find  $Prob(\text{within } t_{sus}\sigma)$ , which is the probability to locate  $x_{sus}$  inside of  $t_{sus}$  times the standard deviation of the distribution. With this, the probability outside of the range given by  $t_{sus}$  times the standard deviation can be calculate as:

$$Prob(\text{outside } t_{sus}\sigma) = 1 - Prob(\text{within } t_{sus}\sigma)$$

which determines the probability that a legitimate measurement would differ from  $\hat{x}$  by  $t_{sus}$  or more standard deviations. Thus, for a  $t_{sus} = 2,45$  the probability of  $Prob(\text{outside } 2,45\sigma) = 1 - 0,9857 = 0,0143$ , where the integer and the first decimal of  $t_{sus}$  determine from the first column of table A1 in Appendix A, the row which will intercept the column of the same table defined by the second decimal of  $t_{sus}$ . This interception locates the probability value  $Prob(\text{within } t_{sus}\sigma)$  corresponding to the analyzed  $x_{sus}$ . Finally, multiplying by  $N$ , the total number of measurements, gives:

$$\begin{aligned} n &= (\text{expected number as deviant as } x_{sus}) \\ &= N \times Prob(\text{outside } t_{sus}\sigma) \end{aligned}$$

If this expected number  $n$  is less than one half, then, according to Chauvenet's criterion,  $x_{sus}$  can be rejected. This means:

$$n < 0,5 \quad \text{then } x_{sus} \text{ can be deleted}$$

Table A1 in Appendix A is named *normal table* and it is very common in handling probabilities for normal distributions. Because tables are used for calculating probabilities with variables that follow a normal distribution and since it would be impossible to have a table for each possible normal distribution, this last is transformed in a standard normal distribution, such that its mean ( $\hat{x}$ ) is zero and its standard deviation ( $\sigma$ ) is one, i.e.,  $N(0,1)$ . It is achieved using equation 3, where: 1) a data shift is performed by subtracting  $\hat{x}$  to each one of the data points, so that the data becomes "centered" in zero in the representation of its *density function*; 2) the standard deviation is equal to 1, by dividing the previous result by the standard deviation of the normal distribution. This transformation is called *standardizing or normalizing of the variable* [82], although the term "normalizing" can make reference to different processes, and the data resulting

from this, are known as *z-values*, *z-scores* or *normal scores*, represented in the previous explanation as  $t_{sus}$ . A graphical representation of this process is shown in figure 21.

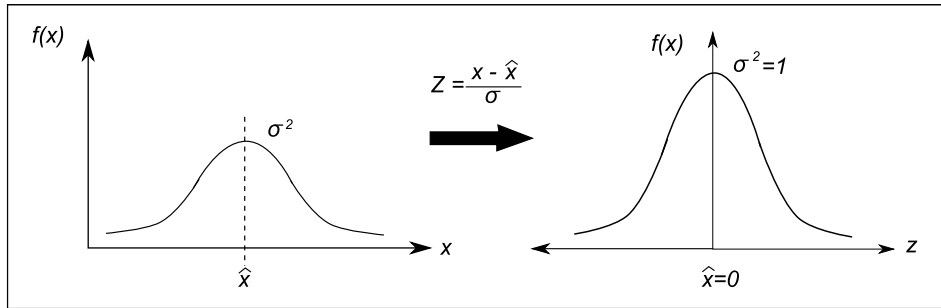


Figure 21. Graphical representation of the standardizing process.

In this way, the Chauvenet criterion is applicable only to normal distributions (as the previous use of the *normal table* suggests it), and although, the irradiance profile and other variables (humidity, wind speed, temperature and sunshine) seem to have a probability distribution similar to a normal distribution, a *normality test* is performed to verify this. The procedure used for the application of this test is explained in Appendix B. Therefore, each one of the data points is evaluated for determining if the set where these belong, represents a normal distribution or not for the application of the Chauvenet’s criterion. A more detailed information about the *standard normal distribution* and the *t-test* (including the characteristics of the *t-distribution*) can be found in [82] and [83].

Considering the previous analysis, an algorithm to apply the Chauvenet’s criterion was developed and its flowchart is presented in figure 22, based on the following procedure where the evaluation of the normality is included:

- The data obtained after the application of the standardized residual method (as in previous analysis in sets of 12 values by day) are verified to determine if its probability distribution is a normal distribution for the evaluation of the Chauvenet’s criterion.
- If the previous step is validated, the normal distribution is transformed in a standard normal distribution, calculating the probability of occurrence of each value outside of its z-score. After this, by multiplying this probability with the number of data points in the set, the Chauvenet’s criterion is evaluated, deleting the value that does not meet the established requirements.
- The deleted data are interpolated or extrapolated according to the position inside of the set and a report about these data is generate for the future review.

At the end, all data sets shown a normal distribution, so the criterion could be applied, and the results of this analysis are presented in table 17 for each one of the stations.

Station	Irradiation	Humidity	Wind speed	Temperature
Universidad Francisco de Paula Santander	0,46 %	2,97 %	1,48 %	3,13 %
Universidad de Pamplona	1,00 %	2,37 %	1,42 %	2,20 %
Alcaldía de Herrán	0,73 %	2,24 %	2,10 %	1,13 %

Table 17. Percentage of modified data by the analysis with Chauvenet’s criterion.

### 3. Daily clearness index ( $K_t$ ) test:

In several works, extraterrestrial solar radiation has been used as a tool in the error detection in the data of the irradiance profile for solar estimation applications [45] [84]. In [53], the authors applied an indirect way using  $K_t$  as an indicator, in order to identify if the experimental values were bigger than the calculated ones from the concept of extraterrestrial solar radiation, which would represent an error.

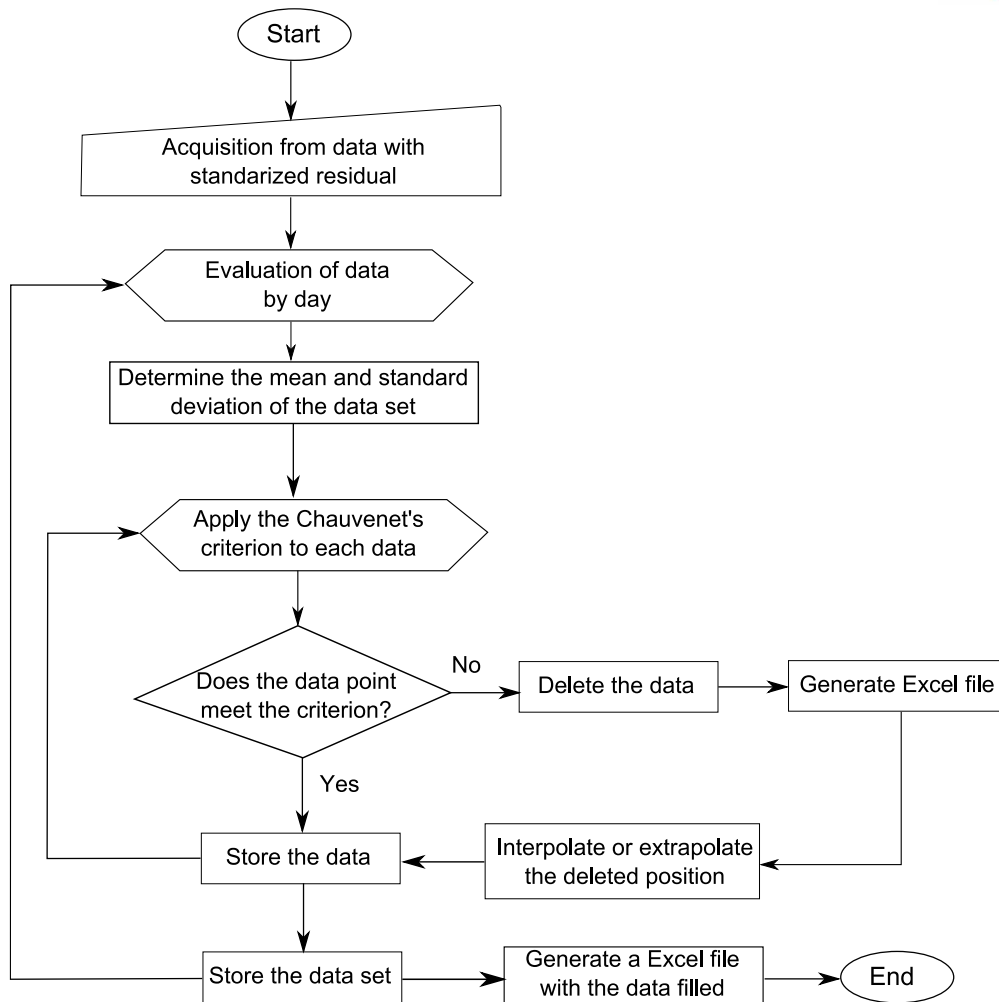


Figure 22. Flowchart for Chauvenet's criterion algorithm in Matlab.

For a better understanding of the above, some concepts are called and described in Appendix C. There, the procedure to calculate the indicator  $K_t$  is detailed to follow its application in the algorithm developed in this analysis.

The algorithm elaborated in Matlab for the execution of this criterion meets the following sequence of steps:

- Data are evaluated by days; from the concept in Appendix C,  $\alpha$  is calculated from the corresponding location of the day under assessment. With this, the declination angle and the incident extraterrestrial solar radiation is determined. Later,  $\omega$  is obtained and finally, the daily extraterrestrial solar radiation is defined with the respective latitude of the place where the data are supplied.
- Average daily solar radiation is calculated from the experimental values for the days analyzed, and together to the results of the previous point, the daily clearness index is established, deleting the data set (data in a day) whose  $K_t$  is outside of the range mentioned at the beginning of the section *step 3*. A report about it is generated by the algorithm.
- The procedure is repeated for all days of the year. Note that the interpolation process is not applied here, because the elimination of all data of the day is performed.

The flowchart of this algorithm developed in Matlab is similar to the one used in the Chauvenet's criterion just that applying the range for the  $K_t$  indicator. The amount of data with errors for all cases, i.e., for all years in each one of the stations did not surpass 1 % of the total data, which were deleted before continuing with the error detection criteria by hour of the irradiation data.

- Step 4: Integration of data

After organizing the data, to detect and modify the atypical values in their structure, an algorithm to integrate the information of the different variables by station was developed. This took each variable and deleted the data whose hours and days were absent in the other variables for each year and station. At the end, the same amount of data for each variable is obtained, allowing to mix all data in the same file.

The amount of data resulting from the integration process is presented in table 18. These data are normalized in the next step to start the training and validation of the ANN. From table 18 can be deduced that the ANN has a total of 52.740 data points by each one of the evaluated variables for its training, validation and testing process. A total of 263.700 data points are processed to construct the estimation model in the present work.

Year	Universidad Francisco de Paula Santander station	Universidad de Pamplona station	Alcaldía de Herrán station
2006	2.568	NS	300
2007	3.012	NS	2.100
2008	3.048	NS	1.020
2009	NS	NS	1.212
2010	1.776	NS	576
2011	3.264	636	84
2012	2.340	108	816
2013	1.368	3.792	1.440
2014	2.700	2.100	2.520
2015	3.156	864	1.368
2016	NS	4.104	4.104
2017	NS	1.236	1.128
<b>Total</b>	<b>23.232</b>	<b>12.840</b>	<b>16.668</b>

NS=Not supplied.

Table 18. Final amount of data points for the normalization process.

- Step 5: Normalization

The normalization process is a common method used in the pre-processing of data, and it consists of handling the input data in a smaller range from a wider one. The normalization range can vary since it must be specified according to the activation function that will be used in the network [28]. Then the appropriate input variables are determined to ensure optimum accuracy of the model. According to [16], the data can be restricted within a range between 0 and 1 to minimize the regression error, improve precision, and maintain correlation among the dataset. In addition to this, the activation functions deployed in Matlab for the training of the AI models mentioned at the end of the section 3.1.1 are designed for ranges between -1 and +1 (except the *purelin* function) so that a range from 0 to 1 is viable for its application. Therefore, this is the first applied normalization range to the data; its implementation is defined by the following formula [4] [16]:

$$X_{norm} = \frac{X_{actual} - X_{min}}{X_{max} - X_{min}} \quad Eq. 4$$

where  $X_{norm}$  is the normalized data;  $X_{actual}$  is the original input; and  $X_{max}$  and  $X_{min}$  are the minimum and maximum values of the input data, respectively.

If the data are pre-processed, then post-processing is required ahead of the calculation or analysis on the performance of the forecasting model. In PV power forecasting, the most commonly used post-processing techniques are anti-normalization and wavelet reconstruction

[16]. If normalized data are used in the forecasting model, then the forecasted value should be anti-normalized to extract the actual forecasting PV power and analyze the performance of the model. Meanwhile, wavelet reconstruction is adopted to extract the actual forecasting PV power if the model input data are pre-processed by wavelet decomposition (WD).

Other expressions can be found in the literature [36] [29] for applying different normalization ranges, being the range from -1 to +1 one of the most common (due to the coupling with the ranges of the activations functions) after the one mentioned from equation 4. It is from the above reason that this range is also used, and its results were compared with the ones obtained from equation 4, in order to select the best performance. The formula for the application of this range is presented in equation 5.

$$X_{norm} = (Max - Min) \left( \frac{X_{actual} - X_{min}}{X_{max} - X_{min}} \right) - Min \quad Eq.5$$

where *Max* and *Min* represent the top and bottom limits of the normalization range, i.e., +1 and -1, respectively for this case.

A simple algorithm of conversion was implemented in this step to transform both the input and output data used in the construction of the AI models.

#### 4.2.2. Design stage: Structure of the network

After the pre-processing stage, the network is ready to train the connections between the climatological variables (inputs) and the irradiation estimation (output) for both, the ANN and the ANFIS models. For this, the present section deals with the following procedure for the two AI techniques: first, different structures and learning techniques are used with the four climatological variables (assuming that their participation in group will show the highest accuracy of the model) for the Universidad Francisco de Paula Santander station in order to find the topology with the best performance. Subsequently, a brief analysis of the input data is performed to examine its influence and participation in the model. After this, the best results from the previous analysis, are extended for the implementation of the ANN and ANFIS models of the remaining two stations in evaluation (Universidad de Pamplona station and Alcaldía de Herrán station), completing the three models expected for the municipalities assessed in this study. Finally, an error comparison is carried with other works in the literature to estimate the behavior of the models. From here on, the municipalities will be called cities for the reader to have a better understanding.

##### - Creation of the ANN structure:

The ANN was developed in Matlab through its *Neural Network Toolbox* which provides a framework for designing and implementing deep neural networks with algorithms, pre-trained models, and apps. It is widely used because it allows to perform tasks of classification, regression, clustering, dimensionality reduction, time-series forecasting, and dynamic system modeling and control in a fast and effective way.

For the creation of the network, the Neural Network Toolbox requests the following parameters: topology, training algorithm, training and target data, number of neurons and layers, type of activation function in each layer and the performance indicator to evaluate the training of the network. The different options that Matlab offers are implemented, taking as initialization point those parameters which showed better results in the literature for solar estimation models:

1. Topology: Feed forward back propagation [4] [28] [29] [53] [85].
2. Training algorithm: Levenberg–Marquardt back-propagation [4] [28] [29] [53] [54] [85].
3. Number of hidden layer: 1 [4] [29] [53].
4. Number of neurons: In [4] a range from 3 to 20 neurons was evaluated with a final value of 10 for the best performance. In [29] the evaluated range was between 1 and 30 neurons, obtaining

the best result with a hidden layer of 24 neurons; and in [53] a range until 80 neurons was implemented with similar results for values between 10 and 20 units. Since for this parameter, there is not a fixed value to show a globalized result, a range between 5 and 30 neurons was applied, considering that the most promising results were found in this interval.

5. Activation function: The hyperbolic tangent sigmoid function and linear function are generally used in the hidden and output layer respectively, so that these were the ones used for the first construction of the network [29] [53]. Other options were also evaluated, for instance, the hyperbolic tangent sigmoid function for both layers (hidden and output, the input layer does not use any activation function since it does not perform any processing data, as it was mentioned in previous sections) which presented the best results in [4] and [85].
6. Performance indicator: In the Neural Network Tool interface, three error indicators can be selected to analyze the performance of the network: MSE, SSE (mentioned in previous sections) and MSEREG which measures network performance as the weight sum of two factors: the mean squared error and the mean squared weight and bias values. The first two are the most common, but in order to be able to compare the results of the model with those in the literature, MSE was selected as performance target. This parameter is very important because the ANN will try to reach that value in the training process; if the MSE value is not achieved after repeated attempts for convergence, other parameters as the number of epochs or validation checks are analyzed for the finalization of the training process. These characteristics are explained in the next sections.

The initial input data used for the training were obtained by the normalization algorithm for the variables of humidity, temperature, wind speed, and sunshine; the target data was the corresponding irradiation values for each combination.

One of the problems that occur during neural network training is the presence of an effect called *overfitting*. The error on the training set is driven to a very small value, but when new data is presented to the network the error is large. The network has memorized the training examples, but it has not learned to generalize to new situations. For this reason, there are two methods for improving generalization that are implemented in Neural Network Toolbox™ software: *regularization* and *early stopping*. The second one is the default method for improving generalization; this technique is automatically provided for all of the supervised network creation functions, including the backpropagation network [86].

In this technique the available data is divided into three groups. The first group is the training set, which is used for computing the gradient and updating the network weights and biases. The second one is the validation set. The error on the validation set is monitored during the training process. The validation error normally decreases during the initial phase of training, and so does the training set error. However, when the network begins to overfit the data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum of the validation error are returned.

The final group is the test set error. It is not used during training, but it is used to compare different models. It is also useful to plot the test set error during the training process. If the error in the test set reaches a minimum at a significantly different iteration number than the validation set error, this might indicate a poor division of the data set.

From the total data points, 80 %, i.e., 42.192 of the 52.740 data for each variable, were used for the training, validation and testing of the network, as it was done in [28] and [29], and the remaining 20 % for the simulation process. The data used for the training, validation and testing were taken randomly of the 80 % of the total data mentioned by the *dividerand* function with a relationship of 70:15:15, respectively, which can be varied according to the needs of the ANN.



The simulation process is applied with different data than the data used in the training, validation and testing processes, and it is implemented as an indicator of the network performance when data are unknown for the model. Here, the output must be analyzed in an independent way of the Neural Network Tool, since it does not analyze the output data in this mode (simulation).

To start the training process, it is needed to establish the stop and run parameters. The first are the characteristics which lead the execution of the learning algorithm and indicate when it must stop. As it was mentioned previously, the main objective for the training process is to reach the error defined by the indicator selected in the creation of the network, that for this case was the MSE; but when the algorithm detects that the error does not decrease considerably with each epoch, and that this error is the minimum which the network with the current characteristics can achieve, then the algorithm uses other parameters to stop the training; some of them are:

- Number of epochs: It represents the maximum number of cycles or epochs that the algorithm will perform.
- Training time: Maximum execution time of the algorithm.
- Minimum performance gradient: It indicates the performance of the gradient, and since this determines the changes of the weights and biases, a small value will indicate that the network is not applying significant changes in the updating of the weights and that is near to reach its goal.
- Maximum validation failures: Maximum number of iterations when the error starts to increase after rising a minimum.

On the other hand, the run parameters define the form in which the training is executed and depend on the training algorithm to implement. For the case of the Levenberg–Marquardt back-propagation algorithm, the following parameters are requested:  $\mu$ ,  $\mu_{inc}$ ,  $\mu_{dec}$  and  $\mu_{max}$ , which express the initial factor  $\mu$ , its increase and decrease, and the maximum value that  $\mu$  can rise before stopping the training. The factor  $\mu$  is a value used to calculate the Jacobian  $jx$ , objective of the back-propagation algorithm, with respect to the weight and bias variables  $X$ . Each variable is adjusted according to Levenberg-Marquardt method:

$$jj = jx \times jx \text{ and } je = jx \times E$$

$$dx = -(jj + I \times \mu)/je$$

where  $E$  is all errors and  $I$  is the identity matrix.

The adaptive value  $\mu$  is increased by  $\mu_{inc}$  until the change above results in a reduced performance value. The change is then made to the network and  $\mu$  is decreased by  $\mu_{dec}$ . Training stops when  $\mu$  exceeds  $\mu_{max}$ . A more detailed information about the Levenberg–Marquardt back-propagation algorithm in Matlab can be consulted in [87].

Table 19 shows the training parameters used for the creation of the ANN, based on the criteria presented in [29]. Different trainings were applied, varying the number of neurons as first evaluation parameter, taking into account that there is not a globalized criterion for its selection. The results in terms of performance with respect to the applied number of neurons after the training is shown in tables 20a and 20b, for the data normalized from equation 4 and 5, respectively. In these tables, the two best results are highlighted for visual identification.

Apart of the MSE indicator, table 20 presents the R indicator provided by Matlab after the training; *R Values* measure the correlation between outputs and targets from a linear regression. An R value of 1 means a perfect linear relationship, and 0 a random relationship. For the case of the MSE indicator, lower values are better, and zero value means no error. Thus, the best model is that whose combination presents the lowest MSE with the highest R.

Parameter	Value
Maximum number of epochs to train	1000
Performance goal	0
Maximum validation failures	6
Minimum performance gradient	1e-5
Initial mu	0,001
Mu decrease factor	0,1
Mu increase factor	10
Maximum mu	1e10
Maximum time to train in seconds	Inf

Table 19. Training parameters.

Number of neurons	MSE	R
5	0,0274	0,832
<b>10</b>	<b>0,0260</b>	<b>0,843</b>
15	0,0264	0,842
20	0,0267	0,842
25	0,0263	0,843
<b>30</b>	<b>0,0260</b>	<b>0,843</b>

a.

Number of neurons	MSE	R
5	0,104	0,844
10	0,102	0,846
<b>15</b>	<b>0,101</b>	<b>0,847</b>
20	0,102	0,846
25	0,102	0,846
<b>30</b>	<b>0,102</b>	<b>0,847</b>

b.

Table 20. Performance indicators for each process of construction of the model in terms of the number of neurons in the hidden layer: a. Data normalized from equation 4. b. Data normalized from equation 5.

As it can be observed in table 20, there is not a significant variation in the performance indicators for changes in the number of neurons in the hidden layer with the implementation of the Levenberg–Marquardt back-propagation algorithm. In contrast, the normalization range demonstrated to have a large impact in the results, whereby the range applied from equation 4 is expected to be the one used in the training process. Therefore, the number of neurons was selected according to the lowest computational capacity that this parameter required for a final value of 10 in the first range and 15 for the second one.

In addition to this, other available learning algorithms were implemented; considering the previous analysis, the same variations in the number of neurons were applied. These as in the Levenberg–Marquardt back-propagation algorithm did not expose large effects in the performance of the training. The best result for each algorithm in terms of the number of neurons is shown in table 21 for the two different normalizations, in order to compare the performance among the learning strategies.

Although the results presented in table 21 are very similar, it verifies that the best performance is obtained with the Levenberg–Marquardt back-propagation, following the conclusions in other research studies. It is worth to mention that all algorithms finished its training process after reaching the number of validation checks in less than 1000 epochs except the Gradient Descent with Momentum algorithm which needed about 2500 epochs to obtain a performance close to the other ones. This indicates that the Gradient Descent with Momentum algorithm was about two times slower in its execution respect to rest of the algorithms.

In this way, the Levenberg–Marquardt back-propagation algorithm is confirmed as learning strategy of the ANN. The results shown in table 21 were obtained using the hyperbolic tangent sigmoid function (*tansig*) as activation function in the hidden layer, and the linear function (*purelin*) in the output layer according to the analyzed literature. Therefore, to identify if this is the best combination, the different activation functions offered by Matlab were applied, generating the results displayed in tables 22a and 22b. These results correspond to the best parameters

obtained until this point for each type of applied normalization, i.e., a feed-forward topology with the Levenberg–Marquardt back-propagation algorithm and one hidden layer of 10 neurons for table 22a, and a similar structure but with 15 neurons for the table 22b.

Algorithm	Number of neurons	MSE	R
Levenberg–Marquardt back-propagation	10	0,0260	0,843
Gradient Descent back-propagation algorithm	25	0,0314	0,815
Gradient Descent with Momentum	5	0,0311	0,815
Resilience back-propagation	20	0,0272	0,838
Scaled conjugate Gradient	20	0,0279	0,835
Conjugate Gradient back-propagation with Fletcher-Reeves Updates	25	0,0293	0,828
Conjugate Gradient back-propagation with Polak-Riebre Updates	25	0,0281	0,834
Broyden-Fletcher Goldfarb Shanno	15	0,0282	0,834

a.

Algorithm	Number of neurons	MSE	R
Levenberg–Marquardt back-propagation	15	0,101	0,846
Gradient Descent back-propagation algorithm	25	0,163	0,741
Gradient Descent with Momentum	5	0,112	0,829
Resilience back-propagation	20	0,117	0,819
Scaled conjugate Gradient	20	0,111	0,835
Conjugate Gradient back-propagation with Fletcher-Reeves Updates	25	0,107	0,837
Conjugate Gradient back-propagation with Polak-Riebre Updates	25	0,109	0,839
Broyden-Fletcher Goldfarb Shanno	15	0,105	0,838

b.

Table 21. Performance comparison among learning algorithms for a feed-forward back-propagation topology of 3 layers: a. Data normalized from equation 4. b. Data normalized from equation 5.

As with the learning strategy, the combination of the activation functions *tansig-purelin* mentioned in the literature reached the best performance indicator (highlighted in blue in the tables), and therefore, it is used in the implementation of the ANN. The worst performance was obtained for the combination *purelin-logsig* highlighted in red in the tables; it is observed that the presence of the *logsig* function in the output layer seems to be the cause of this behavior, since the worst results in the combinations were generated for this relation, although it is a unipolar function optimized at least for the first normalization range. The network shows good indicators up to here, but parameters as the number of layers and topology were modified to find out if these could offer better benefits to the estimation.

In terms of the number of layers, ANNs of 2 and 3 hidden layers were developed; for the first, different combinations from 5 to 30 neurons between the layers were applied and it was found that the performance indicator did not obtain better results than the ANN with one hidden layer. The MSE range generated for these combinations was between 0,0265 (for the combination 20-20 neurons) and 0,0278. Similarly, for the ANNs with 3 hidden layers, different combinations with the same range of number of neurons were implemented without exceeding the performance indicator obtained with one layer, but they did increase the use of the computational resources in their execution. This verifies the concept exposed in [53] where the authors argue that only one hidden layer is needed for this type of model.

Activation function		MSE	R
Hidden layer	Output layer		
<i>tansig</i>	<i>tansig</i>	0,0261	0,842
<b><i>tansig</i></b>	<b><i>purelin</i></b>	<b>0,0260</b>	<b>0,843</b>
<i>tansig</i>	<i>logsig</i>	0,0813	0,674
<i>purelin</i>	<i>purelin</i>	0,0341	0,786
<i>purelin</i>	<i>tansig</i>	0,0289	0,824
<b><i>purelin</i></b>	<b><i>logsig</i></b>	<b>0,0833</b>	<b>0,632</b>
<i>logsig</i>	<i>logsig</i>	0,0812	0,677
<i>logsig</i>	<i>tansig</i>	0,0269	0,839
<i>logsig</i>	<i>purelin</i>	0,0261	0,842

a.

Activation function		MSE	R
Hidden layer	Output layer		
<i>tansig</i>	<i>tansig</i>	0,104	0,842
<b><i>tansig</i></b>	<b><i>purelin</i></b>	<b>0,105</b>	<b>0,843</b>
<i>tansig</i>	<i>logsig</i>	0,326	0,677
<i>purelin</i>	<i>purelin</i>	0,136	0,786
<i>purelin</i>	<i>tansig</i>	0,115	0,824
<b><i>purelin</i></b>	<b><i>logsig</i></b>	<b>0,333</b>	<b>0,633</b>
<i>logsig</i>	<i>logsig</i>	0,328	0,679
<i>logsig</i>	<i>tansig</i>	0,104	0,839
<i>logsig</i>	<i>purelin</i>	0,106	0,841

b.

Table 22. Comparison in terms of performance for different combination of the activation functions: a. Data normalized from equation 4. b. Data normalized from equation 5.

For the analysis of the topology, Cascade-Forward Back Propagation Network (CFB) and Radial Basis Neural Network (RB) were evaluated taking into account that these topologies have been used in the literature for solar estimation [54] [88] [89].

Considering that the data normalized in a range between 0 and 1 has shown the best results in the different analysis, its application was the only one used for continuing with the evaluation based on the topologies; the comparative is displayed in table 23, where it can be concluded that the feed-forward back propagation topology is the best choice for the model. In table 23, Elman Back Propagation network does not show the value for the indicator R because Matlab only use the MSE indicator for the training in this topology. Thus, the final characteristics of the ANN are indicated in table 24. The execution time for all trainings were between 1 and 4 minutes.

Topology	MSE	R
Feed forward back propagation	0,0260	0,846
Cascade-Forward Back Propagation Network	0,0571	0,863
Elman Back Propagation Network	0,0576	-

Table 23. Performance of three topologies for the best behavior of the design parameters evaluated.

Parameter	Value or description
Topology	Feed forward back propagation
Training algorithm	Levenberg–Marquardt back-propagation
Number: of hidden layer/of neurons	1/10
Activation function combination	tansig-purelin (hidden-output layer)

Table 24. Final characteristics of the implemented ANN.

- Analysis of the ANN input data:

After defining the structure of the ANN, the contribution of the input data to the model performance is analyzed. The input data are composed by humidity, temperature, wind speed and sunshine values where each one of these has a participation in the model accuracy depending on its influence and correlation on the solar radiation. For this analysis, the Spearman's rank correlation coefficients [20] are used to determine the correlation among variables. The definition and calculation of Spearman's rank correlation coefficient is as follows:

$$\rho_{x,y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}} \quad Eq.6$$

where  $N$  denotes the number of samples,  $x_i = \{x_1, \dots, x_N\}$  and  $y_i = \{y_1, \dots, y_N\}$  are the variables to relate,  $\bar{x}$  and  $\bar{y}$  are mean values of the two variables, and  $\rho$  denotes the usual Pearson correlation coefficient utilized to rank the variables. Only if all  $N$  ranks are distinct integers, it can be calculated by using:

$$r_s = 1 - \frac{6 \sum_{i=1}^N d_i}{N(N^2 - 1)} \quad Eq.7$$

where  $d_i$  is the difference between two ranks of each sample. If  $|r_s| > 0,8$ , the two variables have a strong correlation, if  $0,2 < |r_s| < 0,8$ , the two variables have a moderate correlation, and if the calculations show a  $|r_s| < 0,2$ , there is a weak correlation between the two variables.

Therefore, the Spearman's rank correlation coefficients among variables and the irradiation are presented in table 25, based on the previous equations.

<b>Spearman</b>	Humidity	Temperature	Wind Speed	Sunshine	Irradiation
Humidity	1	-0,8826	-0,6701	-0,4146	-0,5890
Temperature	-0,8826	1	0,6130	0,4901	0,8206
Wind speed	-0,6701	0,6130	1	0,4477	0,5641
Sunshine	-0,4146	0,4901	0,4477	1	0,8225
Irradiation	-0,5890	0,7206	0,5641	0,7225	1

Table 25. Correlation among climate variables and solar radiation based on Spearman's rank coefficients.

In this way, it can be observed that there is a strong correlation between the temperature and sunshine with the solar radiation, a weak correlation of this variable with the humidity, and a moderate relationship with the wind speed. Then, considering the ANN with the characteristics showed in table 24, several combinations of the input variables were implemented to identify its effect on the performance of the estimation and corroborate the behavior shown in table 25. Table 26 presents the variation of the MSE and R indicators for each combination of inputs.

<b>Variables</b>	<b>MSE</b>	<b>R</b>
H,T,W,S	0,0257	0,843
T,W,S	0,0279	0,831
H,W,S	0,0341	0,786
H,T,S	0,0285	0,828
H,T,W	0,0348	0,784
T,S	0,0298	0,815

Where: H=Humidity, T=Temperature, W=Wind speed and S=Sunshine.

Table 26. Performance of the ANN for different combination of the input variables.

The results observed in table 26 verify the assumptions applied at the beginning of the ANN analysis, where it was argued that the participation of all the variables should generate the best model; from this, the individual impact of each variable was evaluated by its absence in the training of the network, verifying the correlation obtained in table 25, except for the last combination which removes the participation of humidity and wind speed variables of the training.

The results in table 26 express that the variables with the largest influence are the temperature and the sunshine, while humidity and wind speed did not generate a significant impact, although wind speed is slightly more correlated than humidity. Due to the low individual effect of humidity and wind speed, the absence of these variables in the input data of the network could be taken into account for places that do not contain them. Despite of this, the results in table 26 shows that the contribution of these variables is additive whereby if these are available, they can represent a factor to increase (slightly) the model accuracy.

Following the analysis presented in [4], the final allocation of weights from the input layer to the hidden one conducted by the Levenberg-Marquardt training algorithm is presented in table 27, with the corresponding bias value for each neuron. The weights between the hidden and output layer are displayed in table 28, with the output bias for the neuron in the output layer of  $Bias_{out} = -0,2098$ .

Neurons	Weights $_{ij}$	Humidity	Temperature	Wind speed	Sunshine	$Bias_{in}$
1	$\omega_{i,1}$	1,6667	1,4161	-0,21971	1,169	-2,4896
2	$\omega_{i,2}$	0,25031	-1,6255	-1,0005	1,5786	-1,9363
3	$\omega_{i,3}$	-0,95382	1,7498	-0,88941	1,1982	1,3831
4	$\omega_{i,4}$	1,4612	-1,6129	0,90166	0,80516	-0,8298
5	$\omega_{i,5}$	1,6384	1,5719	-0,31681	-0,97069	-0,2766
6	$\omega_{i,6}$	-1,6339	-0,45976	1,0414	1,4942	-0,2766
7	$\omega_{i,7}$	-1,6285	1,1113	-1,1426	1,0028	-0,8298
8	$\omega_{i,8}$	0,065555	1,4482	-0,73774	-1,8847	1,3831
9	$\omega_{i,9}$	1,8767	-0,038279	1,5692	0,46069	1,9363
10	$\omega_{i,10}$	-1,0369	1,8928	-1,0044	-0,72889	-2,4896
<b>Average</b>		<b>0,1705745</b>	<b>0,5453661</b>	<b>-0,179891</b>	<b>0,412437</b>	

\*Where  $i = 1$  for humidity,  $i = 2$  for temperature,  $i = 3$  for wind speed,  $i = 4$  for sunshine, and  $j$  is the neuron number.

Table 27. Allocation of weights by the training algorithm.

Neuron	1	2	3	4	5
<b>Weight<math>_{k,g}</math></b>	0,31396	0,79068	-0,99612	0,77937	0,10378
Neuron	6	7	8	9	10
<b>Weight<math>_{k,g}</math></b>	-0,71936	0,10274	-0,7997	-0,90887	0,5864

Table 28. Weights between hidden and output layer for UFPS station.

The analysis of the contribution of each parameter shows that the most positive weight allocated indicates the highest contribution on a particular neuron in the hidden layer. To simplify the analysis, an average was taken for all weights assigned to a particular input parameter. It was observed that the temperature had the most positive weight, indicating the highest contribution. This was followed by sunshine, wind speed and finally, relative humidity. This order of weights is also in agreement with the expected contribution which was determined in the previous analysis in the tables 25 and 26.

In this sense, the final ANN structure for the Universidad Francisco de Paula Santander station is presented in figure 23.

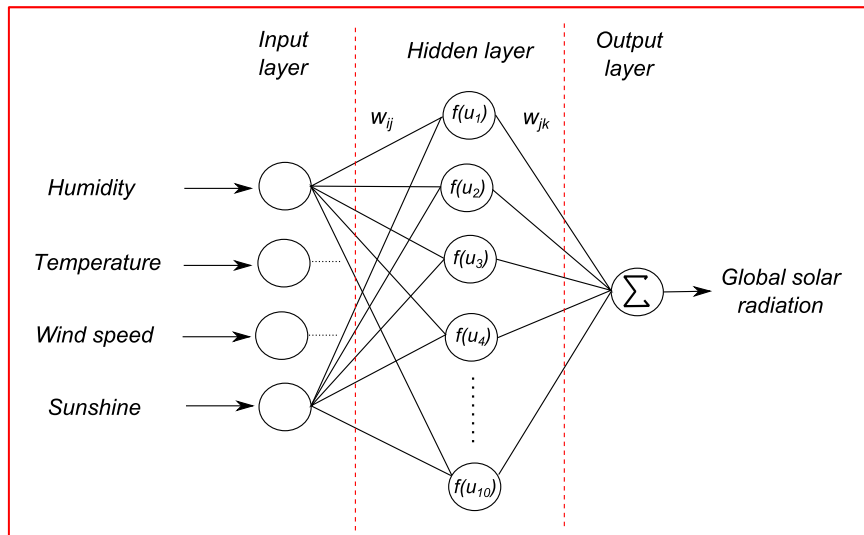


Figure 23. ANN structure with the best performance for the UFPS Station.

The falling curve of the MSE and the evolution of the gradient of the algorithm for the training, validation and testing process are displayed in figures 24a and 24b, respectively. In the first graphic, we can observe that the minimum error is achieved close to the epoch 100. In the second one, it is worth to note that the validation check was executed 6 times before the error started to increase in the epoch 102. To this point the learning algorithm determines that the error does not fall as in the previous validations and therefore, the training process ends indicating that the ANN has reached its best performance with the given characteristics.

The structure in figure 23 is the one defined to globalize and implement in different zones of the region, thus, data from the other stations, i.e., from the *Universidad de Pamplona* and *Alcaldía de Herrán* stations were adapted to construct the estimation model for the locations of Pamplona and Herrán. It means that the structure in figure 23 was trained with the new data to establish the corresponding weights and bias values which could model the irradiation in the aforementioned cities. Table 29 presents the weights obtained for each ANN in Pamplona and Herrán, and table 30 their corresponding weights between the hidden and output layer as in the analysis of the Universidad Francisco de Paula Santander station. The output bias for the last layers for the city of Pamplona and Herrán, were 0,8229 and  $-1,1485$ , respectively.

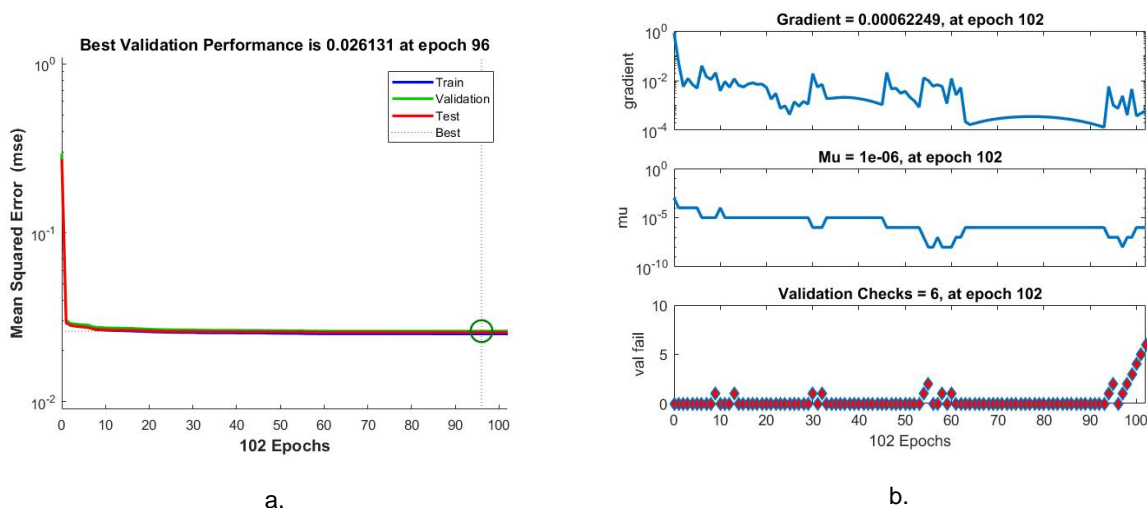


Figure 24. Characteristics of the training process: a. MSE falling curve in the learning process. b. Evolution of the gradient in the Levenberg–Marquardt back-propagation algorithm.

Neurons	Weights	Humidity	Temperature	Wind speed	Sunshine	Bias <sub>in</sub>
1	$\omega_{i,1}$	0,65865	-0,83486	-0,84895	7,5793	8,083
2	$\omega_{i,2}$	0,15098	0,3521	0,30315	1,452	-0,41638
3	$\omega_{i,3}$	-1,0265	3,1569	-0,1542	1,1115	-3,4383
4	$\omega_{i,4}$	2,9196	5,289	-6,0418	-0,95441	-1,847
5	$\omega_{i,5}$	0,63389	2,7232	0,030707	-3,4606	4,8308
6	$\omega_{i,6}$	2,0543	2,2369	-6,2051	-1,424	-2,2048
7	$\omega_{i,7}$	-0,35894	3,0372	1,9327	2,6695	-1,7615
8	$\omega_{i,8}$	2,7585	1,5013	-1,9248	4,0579	4,0292
9	$\omega_{i,9}$	1,7733	0,73603	-1,4597	4,5215	4,6032
10	$\omega_{i,10}$	1,6176	-3,0221	0,36307	-1,2466	3,4822
	<b>Average</b>	<b>1,118138</b>	<b>1,517567</b>	<b>-1,4004</b>	<b>1,4306</b>	

\*Where  $i = 1$  for humidity,  $i = 2$  for temperature,  $i = 3$  for wind speed,  $i = 4$  for sunshine.

a.

Neurons	Weights	Humidity	Temperature	Wind speed	Sunshine	Bias <sub>in</sub>
1	$\omega_{i,1}$	-1,1136	0,93932	-1,0046	-1,56	2,2359
2	$\omega_{i,2}$	-2,1054	1,7954	-2,0379	-2,8894	3,8792
3	$\omega_{i,3}$	-0,54881	-1,1614	0,19828	1,9224	-1,0858
4	$\omega_{i,4}$	-0,08757	-0,31474	-1,7954	-1,8741	1,9645
5	$\omega_{i,5}$	-4,5443	-1,6274	1,3805	3,7179	-4,6142
6	$\omega_{i,6}$	0,34143	2,4075	0,8928	9,9624	-10,1941
7	$\omega_{i,7}$	0,52436	0,44175	-2,0191	-1,8841	2,1724
8	$\omega_{i,8}$	-7,0463	2,3139	0,11936	4,4521	-6,7346
9	$\omega_{i,9}$	-4,4612	3,3916	2,8839	3,1745	-3,9029
10	$\omega_{i,10}$	-0,061374	0,19098	-0,19747	-0,94056	1,1067
	<b>Average</b>	<b>-1,9102764</b>	<b>0,8377</b>	<b>-0,157963</b>	<b>1,408114</b>	

\*Where  $i = 1$  for humidity,  $i = 2$  for temperature,  $i = 3$  for wind speed,  $i = 4$  for sunshine.

b.

Table 29. Weights from the training of the ANN: a. For the city of Pamplona. b. For the city of Herrán.

Neuron	1	2	3	4	5
Weight <sub>k,g</sub>	-2,1759	-0,087314	-1,2434	0,09966	-1,0924
Neuron	6	7	8	9	10
Weight <sub>k,g</sub>	-0,10544	0,042353	-1,7051	2,9407	-0,91491

a.

Neuron	1	2	3	4	5
Weight <sub>k,g</sub>	4,6683	-2,1787	-0,21685	-2,5345	-0,80084
Neuron	6	7	8	9	10
Weight <sub>k,g</sub>	0,71794	1,9092	0,40604	0,34688	-1,3032

b.

Table 30. Weights between hidden and output layer for ANN developed for: a. Pamplona. b. Herrán.

- ANN error analysis:

In order to compare the results of the ANN in figure 23 with other models in the literature, different error indicators were calculated: RMSE, NRMSE, MAE, MAPE and R<sup>2</sup>. Equations from 8 to 12 show the mathematical expressions for each indicator, respectively.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_{measured} - X_{estimated})^2} \quad (Wh/m^2 \text{ or } MJ/m^2) \quad Eq. 8$$



$$NRMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_{measured-N} - X_{estimated-N})^2} \quad (\text{Dimensionless}) \quad \text{Eq. 9}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |X_{measured} - X_{estimated}| \quad (\text{Wh/m}^2 \text{ or MJ/m}^2) \quad \text{Eq. 10}$$

$$MAPE = \left( \frac{1}{N} \sum_{i=1}^N \left| \frac{X_{measured} - X_{estimated}}{X_{measured}} \right| \right) \times 100 \quad (\%) \quad \text{Eq. 11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (X_{measured} - X_{estimated})^2}{\sum_{i=1}^N (X_{estimated})^2} \quad (\text{Dimensionless}) \quad \text{Eq. 12}$$

The MAE is defined as a quantity which is used to measure how close the predicted values are with measured values. The RMSE indicates the level of scatter that the ANN model produces. Lower RMSE indicates that the developed ANN model is having good prediction accuracy [29].  $R^2$  in a similar way to the  $R$  indicator, shows the correlation between the expected and estimated data. Larger  $R^2$  values indicate a stronger matching of trends in the measured data by the model results [53]. Finally, the MAPE indicator represents the average percentage of the difference between the predicted and real values [4].

The results for each station are presented in table 31. This table indicates the error indicators both training and simulation processes; the latter was implemented with the remaining 20 % of the total data for each station (not used in the training process). Although the original data provided by the IDEAM expresses the irradiation in  $\text{Wh/m}^2$ , table 31 also portrays the data in  $\text{MJ/m}^2$  for comparison purposes, since this unit is also common in the literature. According to the International System of Units the equivalence between these units is  $1 \text{ Wh} = 3,6E - 3 \text{ Mega Joules}$ . A comparison of the measured and estimated data is generated in the figures 25, 26 and 27 for the cities under consideration with four random days in the months of January, April, August, and December of the last years provided by the IDEAM for each station. Since there are no seasons in Colombia, these months were selected trying to include different scenarios throughout the year.

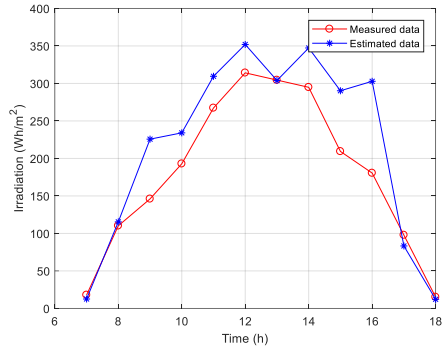
	Cúcuta		Pamplona		Herrán	
	Training	Simulation	Training	Simulation	Training	Simulation
<b>RMSE (Wh/m<sup>2</sup>)</b>	167,47	173,8	113,61	117,6	115,24	116,7
<b>MAE (Wh/m<sup>2</sup>)</b>	121,51	130,2	76,18	80,1	78,9	80,3
<b>MAPE (%)</b>	19,55	21,3	17,08	19,7	17,36	18,4
<b>R<sup>2</sup></b>	0,8941	0,887	0,9139	0,9078	0,9148	0,9101
<b>NRMES</b>	0,1597	0,163	0,1085	0,1107	0,1117	0,1191

a.

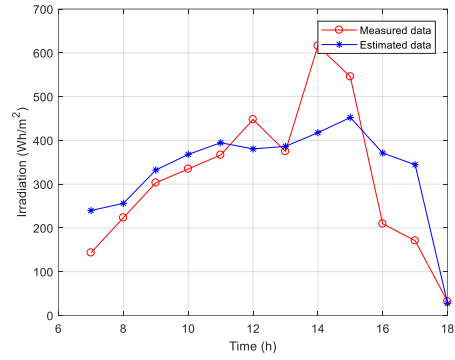
	Cúcuta		Pamplona		Herrán	
	Training	Simulation	Training	Simulation	Training	Simulation
<b>RMSE (MJ/m<sup>2</sup>)</b>	0,60288	0,6256	0,4089	0,4233	0,4148	0,4201
<b>MAE (MJ/m<sup>2</sup>)</b>	0,4374	0,4687	0,2742	0,2883	0,2840	0,289
<b>MAPE (%)</b>	39,55	41,3	27,08	29,7	27,36	28,4

b.

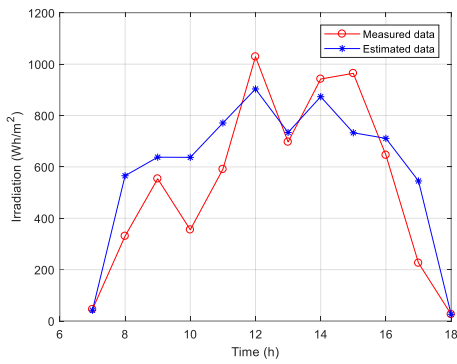
Table 31. Error indicators of ANN model for each one of the evaluated cities in: a.  $\text{Wh/m}^2$ . b.  $\text{MJ/m}^2$ .



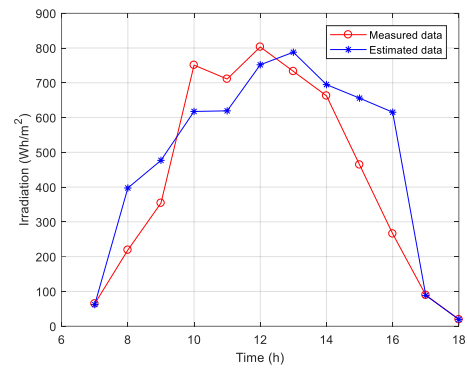
a.



b.

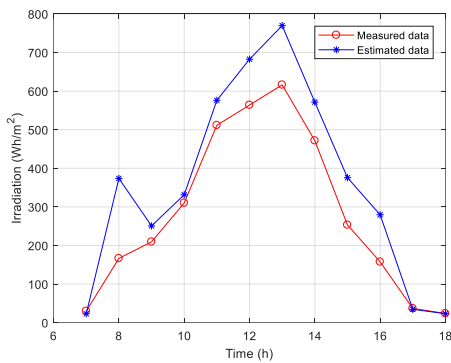


c.

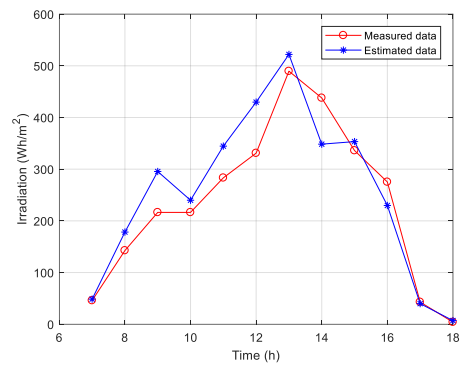


d.

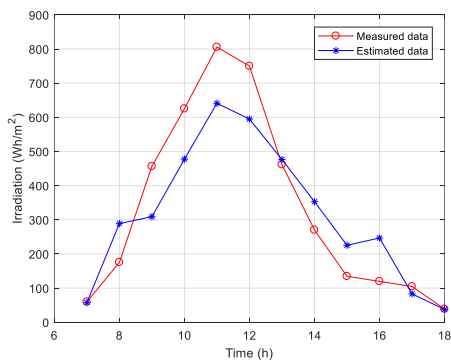
Figure 25. Measured vs. Estimated data for hourly estimation model in the city of Cúcuta: a. January 8, 2015. b. April 6, 2015. c. August 23, 2015. d. December 19, 2015.



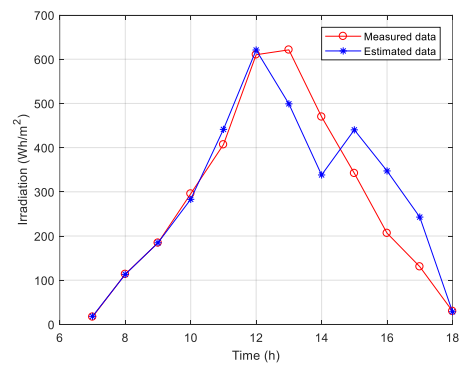
a.



b.



c.



d.

Figure 26. Measured vs. Estimated data for hourly estimation model in the city of Pamplona: a. August 15, 2016. b. December 2, 2016. c. January 6, 2017. d. April 14, 2017.

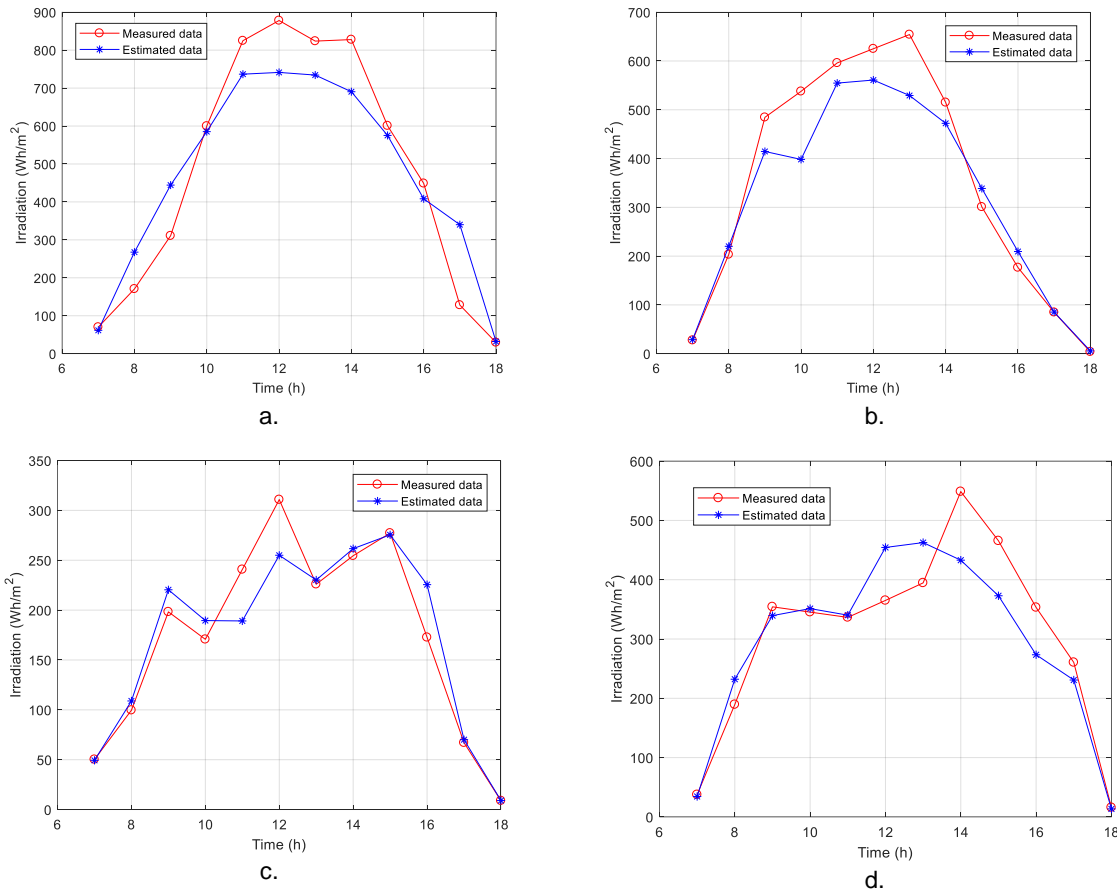


Figure 27. Measured vs. Estimated data for hourly estimation model in the city of Herrán: a. August 6, 2016. b. December 12, 2016. c. January 7, 2017. d. April 21, 2017.

- *Effect of the estimation range:*

Some authors [4] [30] argued to use monthly solar estimation instead of its daily forecasting by the fact that the daily solar radiation series is unstable due to the fast changes in weather conditions. Considering this, it can be expected that the hourly solar estimation presents even greater variations. Despite the above, each time series has its advantages according to the application as it will be exposed in the next sections.

In order to analyze the changes in terms of performance of the mentioned time series, daily and monthly estimation models were trained for each of the stations analyzed in the hourly scale and their results are presented below.

Since the data in previous sections were provided and organized in hourly intervals, an algorithm in Matlab was developed to group them in daily and monthly periods and later, these were normalized applying equation 4. For the monthly case, the same criteria used in the hourly analysis for the amount of missing data was performed. Thus, if a month had more than 5 missing data (close to 16 % of the total data in a month), this month was not taken into account in the evaluation. For 5 or less missing values the interpolation process was applied to fill the data and to obtain the corresponding monthly irradiation data.

Adopting the same ANN structure that the one used in the hourly analysis, the results for daily and monthly estimation are shown in table 32. Unlike table 31 the MSE and R values of the training process are included, and results in the simulation part for the monthly scale are not shown because the amount of data were not enough to support both processes (training and simulation).

		Cúcuta		Pamplona		Herrán	
		Training	Simulation	Training	Simulation	Training	Simulation
Daily	RMSE (Wh/m <sup>2</sup> )	881	959,2	786,22	738,66	636,64	590,03
	MAE (Wh/m <sup>2</sup> )	695,4	760,33	631,96	612,46	498,35	481,1
	NRMES	0,1169	0,1238	0,1259	0,1183	0,1008	0,0935
	MAPE	15,39	17,52	15,8	14,29	13,9	12,86
	R <sup>2</sup>	0,961	0,9398	0,9354	0,937	0,96	0,9627
	MSE	0,0128	-	0,0154	-	0,00957	-
	R	0,775	-	0,796	-	0,836	-
Monthly	RMSE (Wh/m <sup>2</sup> )	3707,4	-	1327,4	-	4651,2	-
	MAE (Wh/m <sup>2</sup> )	2588,8	-	1097,3	-	3443,1	-
	NRMES	0,0842	-	0,0885	-	0,079	-
	MAPE	1,72	-	9,68	-	2,92	-
	R <sup>2</sup>	0,9714	-	0,9852	-	0,98	-
	MSE	0,00115	-	6,3E-11	-	9,28E-06	-
	R	0,947	-	0,792	-	0,947	-

a.

		Cúcuta		Pamplona		Herrán	
		Training	Simulation	Training	Simulation	Training	Simulation
Daily	RMSE (MJ/m <sup>2</sup> )	3,17	3,45	2,83	2,65	2,29	2,12
	MAE (MJ/m <sup>2</sup> )	2,5	2,73	2,27	2,2	1,79	1,73
Monthly	RMSE (MJ/m <sup>2</sup> )	13,34	-	4,77	-	16,74	-
	MAE (MJ/m <sup>2</sup> )	9,31	-	3,95	-	12,39	-

b.

Table 32. Error indicators for daily and monthly estimation models: a. Results in Wh/m<sup>2</sup>. B. Results in MJ/m<sup>2</sup>.

As it can be identified in a comparison between tables 31 and 32, the increase of the estimation period improves the performance of the model; there is an important decrease of indicators as the MAPE and NRMSE, and a growth in the correlation between the measured and estimated irradiation, evaluated by the coefficient of determination  $R^2$ . The MAPE decreased from 19,55 % in the hourly model to 1,72 % for the monthly one in the city of Cúcuta (UFPS station). In the city of Pamplona, it dropped from 19,07 % to 9,68 %, and in the city of Herrán it was reduced from 17,36 % to 2,92 %. In a similar comparison, the coefficient of determination  $R^2$  increased from 0,8941 to 0,9714, from 0,9139 to 0,9852 and from 0,9148 to 0,98, for the cities of Cúcuta, Pamplona and Herrán, respectively. A similar comparison to that in figures 25, 26, and 27 for the hourly estimation is performed in figures 28, 29, and 30 for the daily estimation model for a pair of months with the data provided for last year by the IDEAM which met the minimum amount of missing data allowed for the analysis.

In a similar way to the previous analysis with the daily estimation, the results of the monthly model are portrayed in figures 31a, 31b and 31c for the cities of Cúcuta, Pamplona and Herrán, respectively; in these figures, the months with less than 5 missing values of the last two years for each station were plotted. The good results expressed in table 32 for monthly estimation are confirmed by the closeness between the measured and estimated data in those figures.

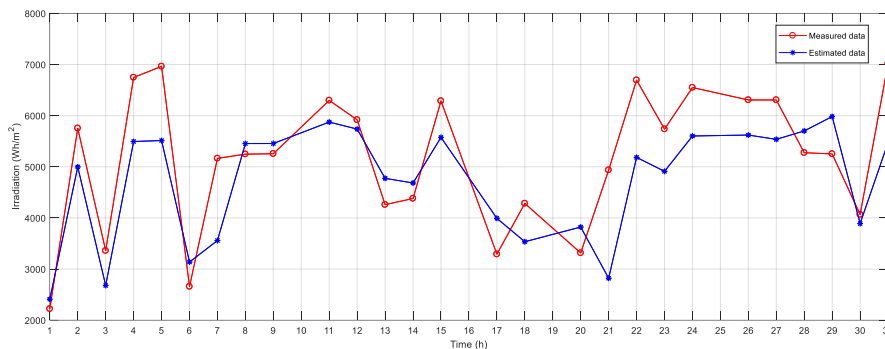
In this study, we consider the MAPE indicator as one of the most accurate for comparison purposes because it is as a normalized error, since it defines the error with respect to the real or measured value, as it is shown in equation 11; meanwhile the other indicators calculate an error

which depend on the range of the variable under analysis. For instance, a RMSE of 200 Wh/m<sup>2</sup> could be a result better than 150 Wh/m<sup>2</sup>, although 200 Wh/m<sup>2</sup> is higher than 150 Wh/m<sup>2</sup>, if the range of the variable for the first one is 1000 Wh/m<sup>2</sup> and for the second one 500 Wh/m<sup>2</sup>. Therefore, the MAPE is used as reference for the evaluation of the models and the rest of error indicators to compare with other works.

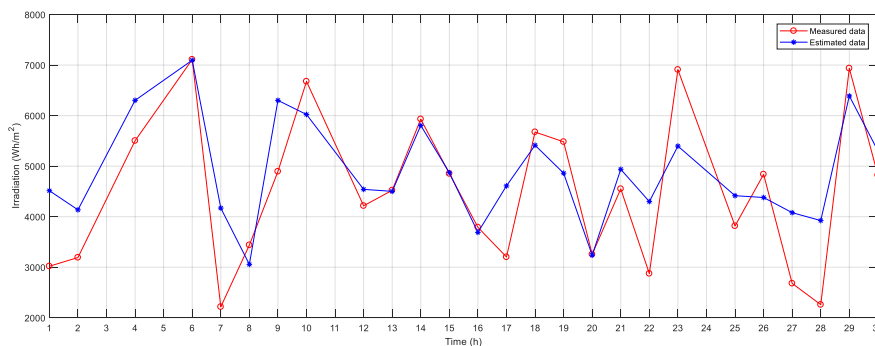
According to [90], in practical applications, MAPE ≤ 10 % means high accuracy, 10 % ≤ MAPE ≤ 20 % means good prediction, 20 % ≤ MAPE ≤ 50 % means reasonable prediction and MAPE ≥ 50 % means inaccurate prediction. Taking the above into account, the hourly and daily models could perform estimations with a good accuracy and the monthly prediction with a high precision, referring to tables 31 and 32.

On the other hand, the results in terms of the error indicators from some works in the literature for the different estimation ranges are presented in table 33, and the research studies where the error is equal or higher than those presented in the current work are highlighted in blue as a performance indicator. This analysis shows that the monthly model has better results than several works in the literature, and that the hourly and daily models record similar conclusions mainly in reference to the R<sup>2</sup> indicator.

One could argue that using the error indicators only to select which model is better than other could be irresponsible considering that each model is trained with very different conditions and objectives. In consequence, the comparison in table 33 is only an overview intended to show what was expected by the models. The real advantages and performance conclusions are evaluated in specific cases for sizing PV systems in the Norte de Santander region in the next sections of the document, where a comparison with similar characteristics is performed.

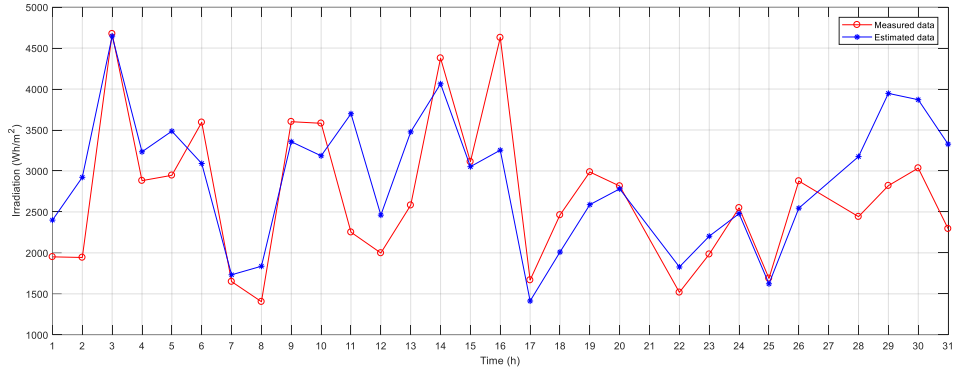


a.

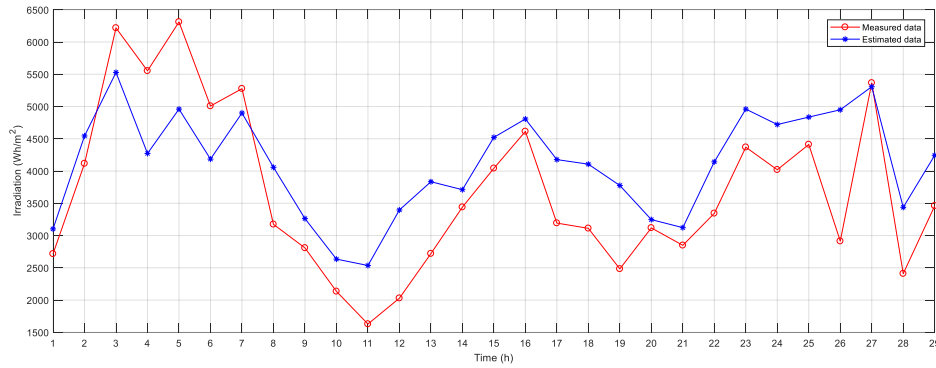


b.

Figure 28. Measured vs. Estimated data for daily estimation model in the city of Cúcuta: a. January, 2015. b. June, 2015.

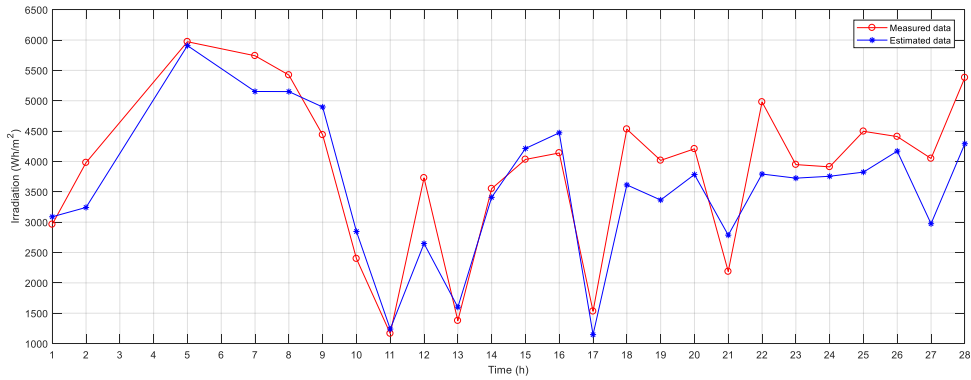


a.

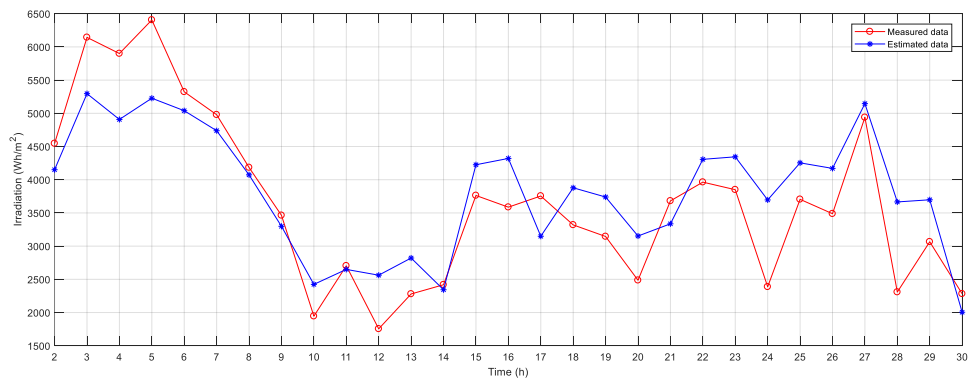


b.

Figure 29. Measured vs. Estimated data for daily estimation model in the city of Pamplona: a. March, 2017. b. April, 2017.

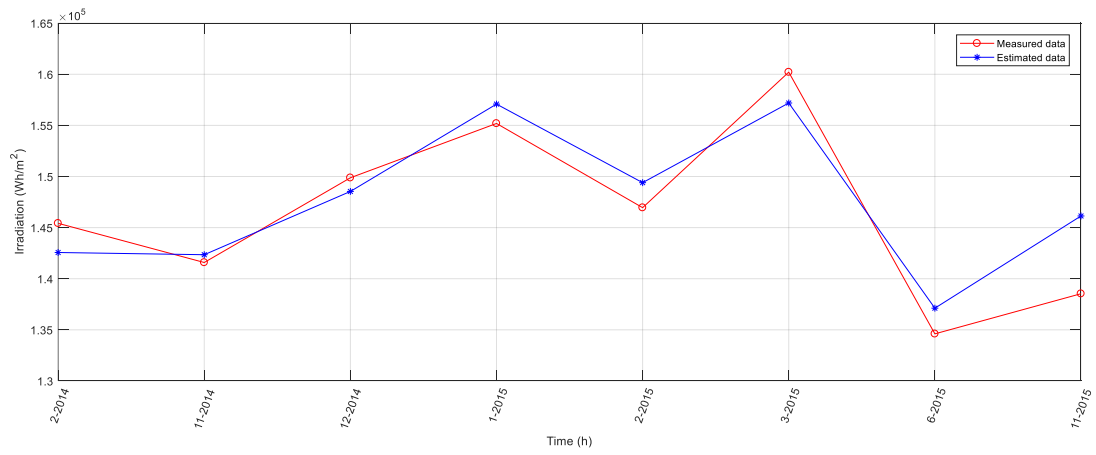


a.

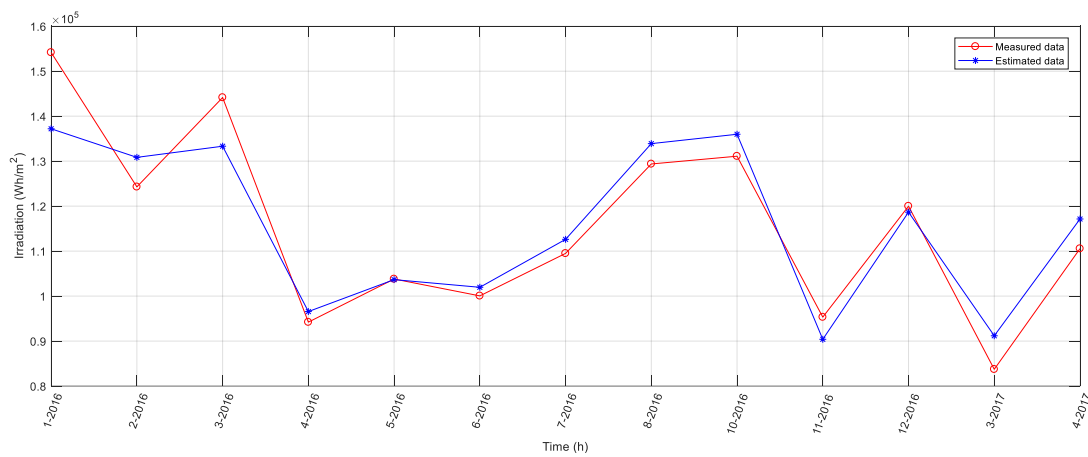


b.

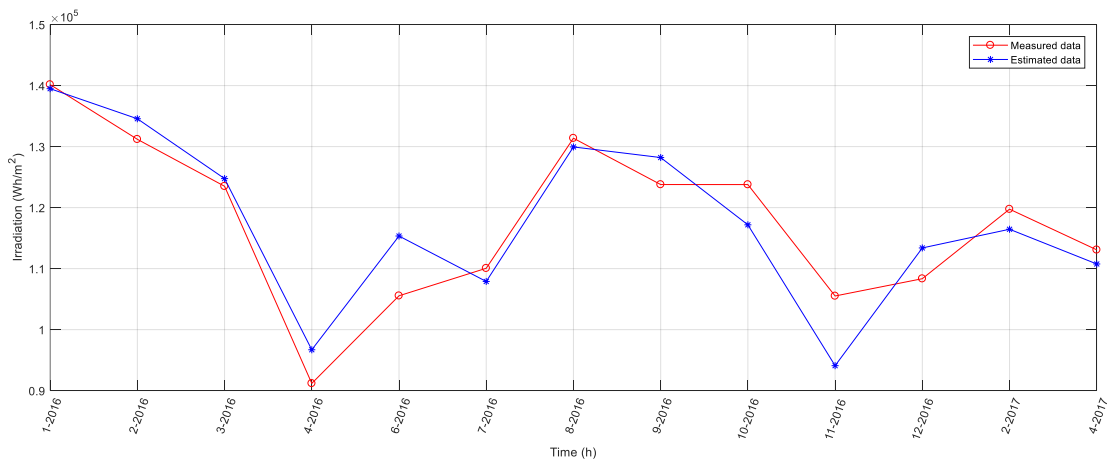
Figure 30. Measured vs. Estimated data for daily estimation model in the city of Herrán: a. February, 2017. b. April, 2017.



a.



b.



c.

Figure 31. Measured vs. Estimated data for monthly estimation model in the city of: a. Cúcuta. b. Pamplona. c. Herrán.

Reference	Input parameters	ANN type	Data recording interval	Error indicator
Ibrahim [48]	Sunshine ratio, humidity, ambient temperature, month and day number, and number of hours by day.	Feed-forward ANN topology	1 year	MAPE(%)= 17,15
				RMSE(%)= 26,44
Chen et al. (2013) in [30]	Temperature, Solar radiation, Sky condition	Feed-forward Neural Network with fuzzy logic	No supplied	MAPE(%)= 6,03
Benmouiza and Cheknane (2013) in [30]	Global horizontal solar radiation time series	Feed-forward Neural Network	1 year	NRMSE= 0,2003
				RMSE(Wh/m <sup>2</sup> )= 64,34
Khosravi [91]	Local time, temperature, pressure, wind speed, and relative humidity	Radial basis function neural network	2010-2016	R(%)= 88,1
		Multilayer feed-forward neural network		RMSE(Wh/m <sup>2</sup> )= 150,141
				R(%)= 98,87
RMSE(Wh/m <sup>2</sup> )= 41,08				
Shaddel [92]	Declination angle, Solar altitude angle, Horizontal global irradiance, Extraterrestrial horizontal solar irradiation	Multilayer perceptron (MLP) using feed-forward back-propagation (BP)	2013	R <sup>2</sup> (%)= 92,42
				MAE= 0,0284
				RMSE= 0,055
Renno [93]	Clearness index (kt), declination angle (d), hour angle (HRA) and global normal irradiance (GNI)	Feed-forward neural network	8 months	MAPE(%)= 5,57
				RMSE(Wh/m <sup>2</sup> )= 17,7
				R <sup>2</sup> (%)= 99,4
Ihya [94]	Clearness index, hour of the day, solar altitude	MLP (Multi-Layers Perceptron) Artificial Neural Network	2009-2011	R(%)= 94
				RRMSE(%)= 20

a.



Reference	Input parameters	ANN type	Data recording interval	Error indicator
Lunche Wang [Ya]	Air temperature, air pressure, relative humidity, water vapor pressure and sunshine duration	ANN multilayer perceptron (MLP), Generalized Regression Neural Network (GRNN), Radial Basis Neural Network (RBNN).	1961-2014	$R^2(\%) = 73-92$
				RMSE(Wh/m <sup>2</sup> )= 538,89 - 908,34
				MAE(Wh/m <sup>2</sup> )= 425 - 636,11
Quej [53]	Minimum and maximum air temperatures, rainfall and global solar radiation.	Feed-forward ANN topology	2000-2014	$R^2(\%) = 62,7 - 68$
				RMSE(Wh/m <sup>2</sup> )= 708,61 – 764,7
				MAE(Wh/m <sup>2</sup> )= 520,28- 602,78
Yildirim [28]	Global solar radiation with Actinographies and sunshine duration	Feed-forward, back-propagation, multilayer perceptron neural network	1997-2007	$R^2(\%) = 96,11$
				RMSE(Wh/m <sup>2</sup> )= 38,88
Bou-Rabee [31]	-	Multilayer feed-forward ANN	2007-2010	MAPE(%)= 86,3
Xue [95]	Month of the year, sunshine duration, mean temperature, rainfall, wind speed, relative humidity, daily global solar radiation Hg.	Back propagation neural network	1995-2014	$R^2(\%) = 94 - 95,7$
				RMSE(Wh/m <sup>2</sup> )= 213,89 - 255,5
				MAE(Wh/m <sup>2</sup> )= 188,33 - 224,44
Zou [96]	Sunshine duration hours, temperature, relative humidity, precipitation, air pressure, water vapor pressure, and wind speed	ANN multilayer perceptron (MLP) with Back Propagation	1994-2010	RMSE(Wh/m <sup>2</sup> )= 302,78 - 736,12
				$R^2(\%) = 80 - 94$
Marzouq [97]	Rainfall, wind direction, daily temperature gradient and global solar irradiation at the top of the atmosphere	Feed-forward Multi-Layer Perceptron (MLP) with evolutionary artificial neural networks (EANN)	2009-2015	$R^2(\%) = 97,5$
				RRMSE(%)= 17,85
				MAPE(%)= 18,46

b.

Reference	Input parameters	ANN type	Data recording interval	Error indicator
Chiteka [4]	Altitude, latitude, longitude, humidity, pressure, clearness index and average temperature	Multilayer feed-forward back propagation	2004-2005	$R^2(\%) = 98,86$
				RMSE(Wh/m <sup>2</sup> )= 223
				MAE(Wh/m <sup>2</sup> )= 170
				MAPE(%)= 2,56
Alsina [27]	Top of Atmosphere (TOA) radiation, day length, number of rainy days and average rainfall, latitude and altitude	Feed-forward Multi-Layer Perceptron (MLP)	1 year	MAPE(%)= 1,67 - 4,25
Neelamegam [29]	Latitude, longitude, altitude, year, month, mean ambient air temperature, mean station level pressure, mean wind speed and mean relative humidity	Feed-forward ANN topology	200-2009	$R^2(\%) = 93,63 - 95,45$
				MAPE(%)= 0,11 - 4,24
				RMSE(Wh/m <sup>2</sup> )= 289,33 - 300,53
Jiang (2009) in [30]	Longitude, Latitude, Altitude, Sunshine percentage	Feed-forward back-propagation	1995-2004	$R^2(\%) = 95$
				RMSE(Wh/m <sup>2</sup> )= 238,89
Ozgoren et al. (2012) in [30]	Latitude, longitude, altitude, Month, mean land surface temperature	Feed-forward back-propagation	2000-2006	MAPE(%)= 5,34
Celik [90]	Monthly mean sunshine duration, monthly mean temperature, altitude, month	Back propagation multilayer perceptron ANN	2000-2010	MAPE(%)= 2,802 - 4,162
				MAPE(%)= 4,082 - 5,27
				MAPE(%)= 3,303 - 6,82
				MAPE(%)= 3,174 - 6,301
Mohammadi et al. (2015) in [14]	Sunshine, temperature	No supplied	No supplied	MAPE(%)= 13,43

c.

Table 33. Error indicator for several models in the literature: a. Hourly. c. Daily. c. Monthly.

- *Creation of the ANFIS structure:*

Using the same data as in the analysis of the ANN model, several ANFIS structures were implemented, considering the three design parameters for this type of networks: the number of membership functions, their type and the learning strategy to apply. Fuzzy Logic Toolbox™ software from Matlab was the tool used for the ANFIS implementation, since it provides command-line functions and an app for training Sugeno-type fuzzy inference systems using specific input/output training data. Unlike the *Neural Network Toolbox™* (which uses the MSE indicator), Fuzzy Logic Toolbox™ uses the RMSE as error indicator for the training, checking and testing processes, so this is the selection parameter of the best structure for the estimation model.

In this type of systems, the expected output is a linear expression as it was indicated in chapter 2; an important step is the initialization of its coefficients to carry out a faster and more accurate training. Thus, a qualitative analysis of the correlation of the input variables with the irradiation is performed to establish a first approximation of the links input-output of the network, and to initialize the coefficients of the ANFIS linear outputs. The results of the analysis were the following:

- It is expected a decrease of irradiation to an increase in humidity values due to the presence of the water vapor particles in the line of sun sight.
- It is expected an increase of the irradiation to an increase of the air temperature.
- It is expected an increase of the irradiation with the increase of the sunshine.
- It is expected a decrease of the temperature with the increase of the wind speed, which decrease the capability of the air to maintain water (decrease the relative humidity) and improves the reception of solar radiation.

It is worth to indicate that the humidity data evaluated in this work are referred to the concept of *relative humidity* for each city; this means that the features of the concept “relative” were taken into account in the previous arguments. A brief explanation of this is given below.

The amount of water vapor particles in the air is known as *humidity*. A mass of air can contain a certain amount of water vapor particles, and this amount depends on the temperature. As the air temperature increases, the air is able to hold more humidity. Therefore, the air has less capacity to contain water vapor at 5 °C than at 15 °C. When a mass of air reaches the maximum amount of water vapor that it can hold at a specific temperature, it is said that the air is *saturated*. Considering this, since the saturation state of the air can vary according to the temperature, the term of *relative humidity* is used to standardize the measurement; thus, the *relative humidity* relates the amount of water vapor present in a mass of air respect to the amount of water vapor which could be present in the same mass of air, if it were totally saturated.

Therefore, an increase in the relative humidity value does not necessarily represent an increase of the humidity or the amount of water vapor particles, if temperature variations are present. In this way, in order to establish first expected conditions of the irradiation with respect to the humidity, it is assumed that the amount of water vapor particles in the evaluation zones is constant, and thus, the changes in the relative humidity values indicate variations in the amount of water vapor particles that the air can group.

Based on all the above, for the number of Membership Functions (MFs) evaluated, a specific number of rules was expected (as it was indicated in chapter 2), and according to the possible combinations of the input data in these MFs, the corresponding coefficients of each linear output of the Sugeno relation *if-then* were predicted. It is expected that this initialization process improves the update of the weights in the training of the ANFIS.

Thus, the number of MFs under evaluation were 2, 3, and 4 and the initial parameters for the mapping of the MF in an input range [0,1] (defined by the normalization range selected) of each variable are shown in table 34, in terms of their most common shapes.

Number of MF by input	<i>i</i>	Triangular			Trapezoidal				Bell-shaped		
		<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>a</i>	<i>b</i>	<i>c</i>
2	1	-1	0	1	-1	-0,3	0,3	1	0,45	2	0
	2	0	1	2	0	0,7	1,3	2	0,45	2	2
3	1	-0,5	0	0,5	-0,5	-0,1	0,1	0,5	0,25	2	0
	2	0	0,5	1	0	0,4	0,6	1	0,25	2	0,5
	3	0,5	1	1,5	0,5	0,9	1,1	1,5	0,25	2	1
4	1	-0,33	0	0,33	-0,33	-0,1	0,1	0,33	0,2	2	0
	2	0	0,33	0,67	0	0,23	0,43	0,67	0,2	2	0,35
	3	0,33	0,67	1	0,33	0,57	0,77	1	0,2	2	0,7
	4	0,67	1	1,33	0,67	0,9	1,1	1,33	0,2	2	1

{*a, b, c*} is the parameter set indicated in figure 10.

Table 34. Parameters for the implemented membership functions.

Figures 32, 33 and 34 show graphs of the triangular, trapezoidal and bell-shaped membership functions used in the training process, respectively, for the parameters exposed in table 34. The performance results for each of the ANFIS structures are presented in table 35. The error indicator applied in this table is the NRMSE, because the data used in the training are the data normalized from the Universidad Francisco de Paula Santander station and the RMSE is the performance indicator in the ANFIS training in Matlab, as it was indicated previously. In the training, two learning strategies were implemented: one based on the back-propagation approach and another on a derived hybrid algorithm.

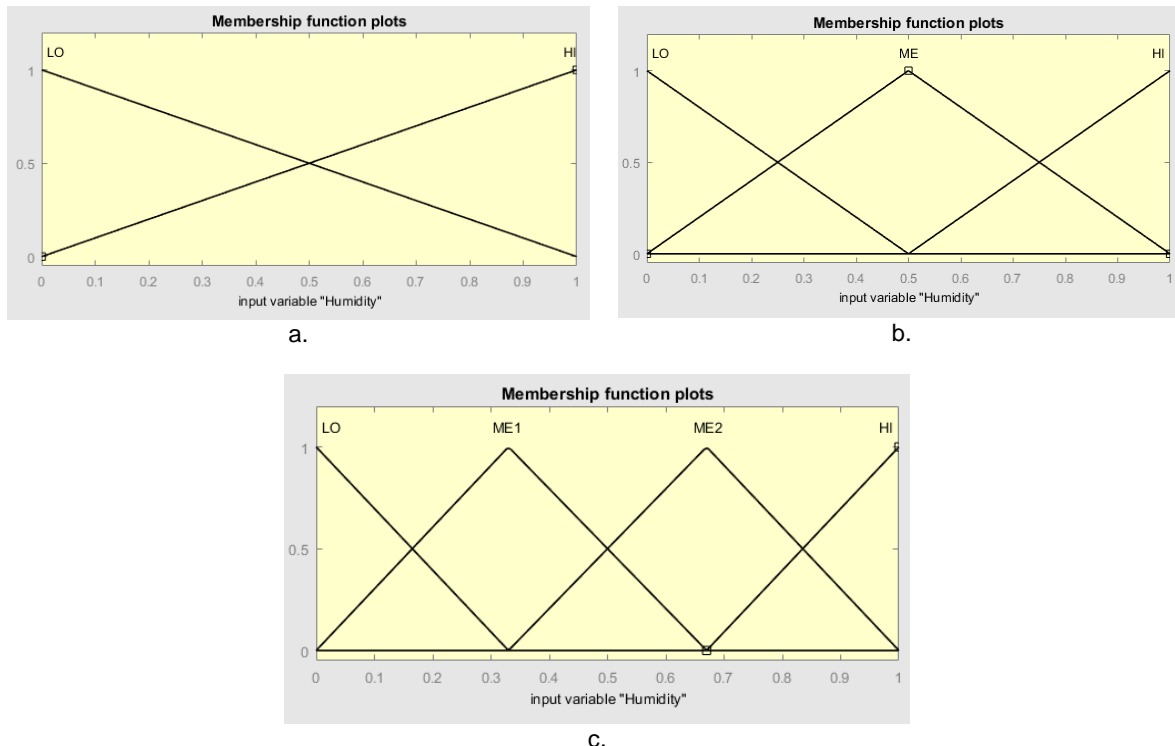


Figure 32. Graphical representation of the membership functions in Matlab: Triangular.

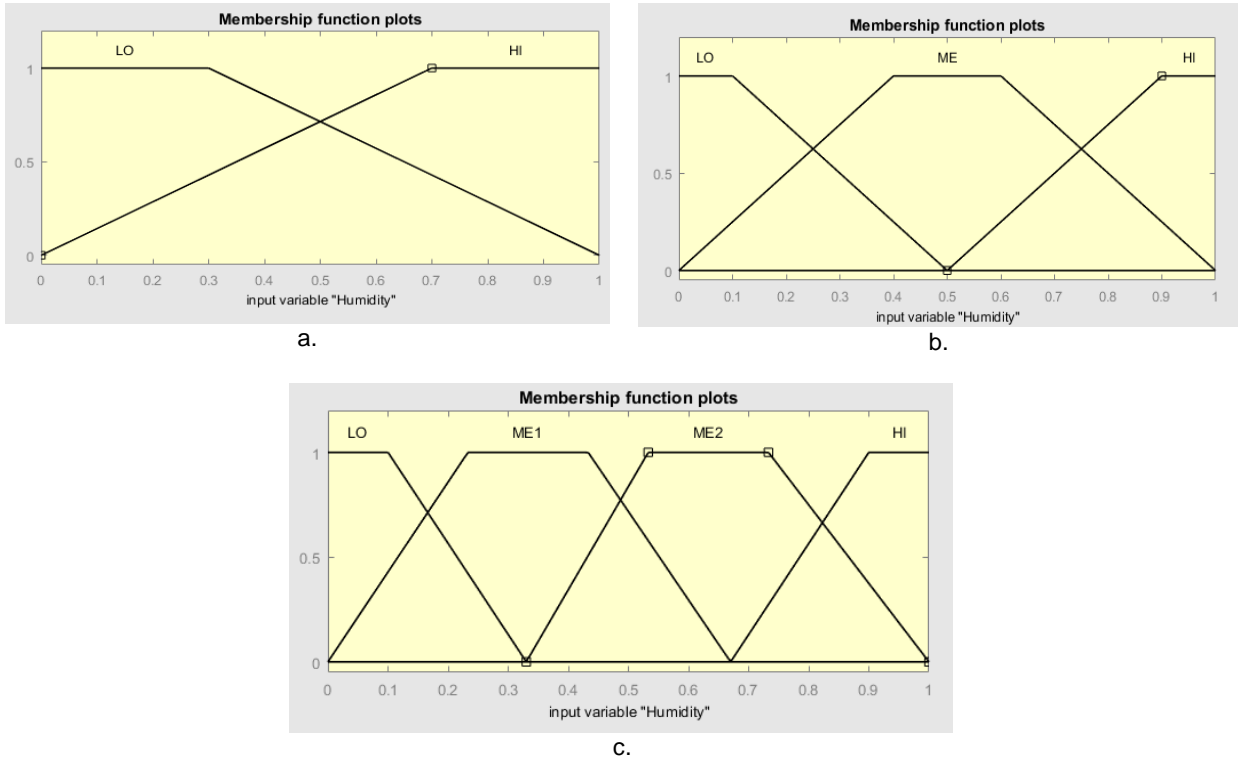


Figure 33. Graphical representation of the membership functions in Matlab: Trapezoidal.

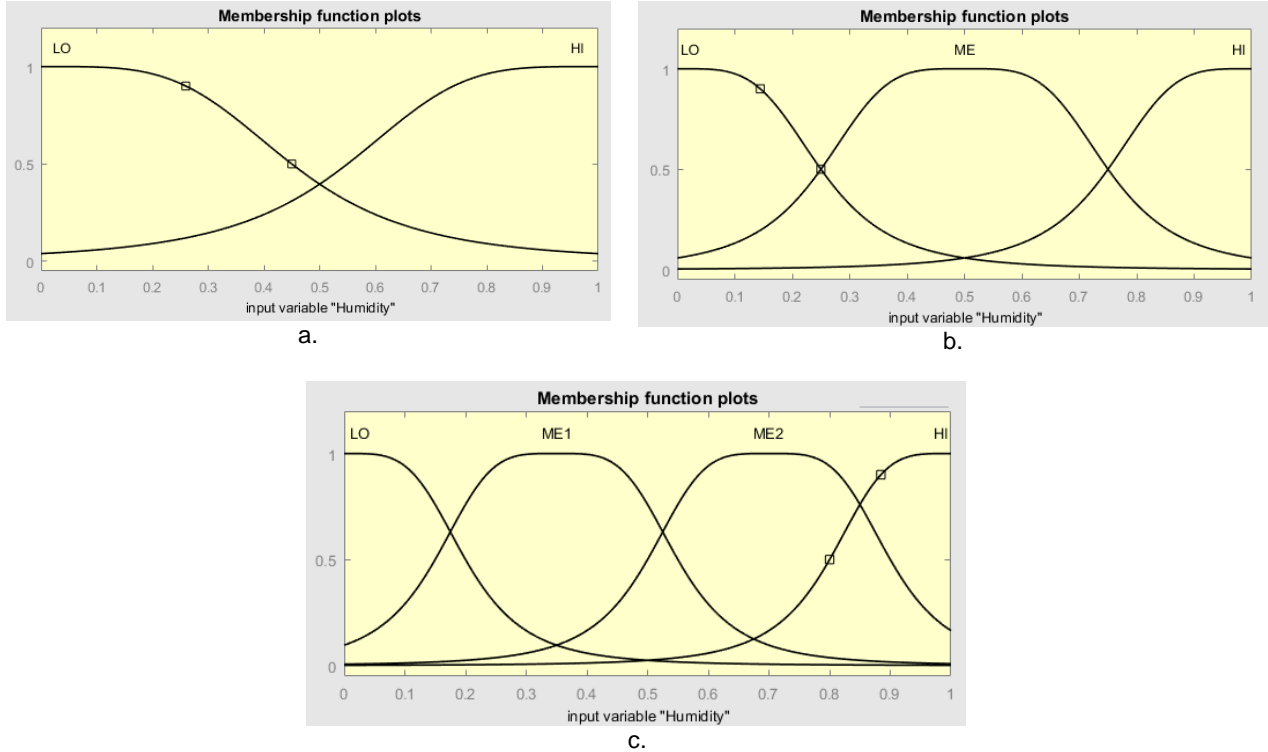


Figure 34. Graphical representation of the membership functions in Matlab: Bell-shaped.

Membership functions	Number of MF	NRMSE	Epochs	Time (min.)
Triangular	2	0,1876	50	12:23
	3	0,1601	30	27:30
	4	0,1592	6	57:14
Trapezoidal	2	0,1620	40	18:21
	3	0,1613	25	29:10
	4	0,1612	4	87:20
Bell-shaped	2	0,1604	60	21:32
	3	0,1602	20	11:02
	4	0,1596	6	59:15

a.

Membership functions	Number of MF	NRMSE	Epochs	Time (min.)
Triangular	2	0,1619	5	1:20
	3	0,1601	5	6:10
	4	0,1592	5	27:05
Trapezoidal	2	0,1620	10	1:40
	3	0,1613	5	5:10
	4	0,1612	5	32:14
Bell-shaped	2	0,1624	10	1:50
	3	0,1602	5	5:50
	4	0,1596	5	29:32

b.

Table 35. Performance of the ANFIS implementations: a. With back-propagation algorithm. b. With hybrid algorithm.

Considering the results in table 35, it can be observed that the number of membership functions does not generate a large impact in the performance of the model but increases the execution time and the computational resources for its implementation. Therefore, an ANFIS structure with two triangular membership functions, linear output relationship and a hybrid optimization method (with better execution time than back-propagation algorithm also in table 35) is enough for the solar estimation model. In table 35 was included the execution time because it varies very much for each implementation unlike of the ANN approach which was very stable and short among changes.

- *Analysis of the ANFIS input data:*

As with the data of the ANN, the influence of each input was analyzed in this section. Figure 35 shows the participation of each variable with respect to the output according with a report of the Fuzzy Logic Toolbox™ from Matlab; humidity has a contribution close to 14 %, the temperature about 26 %, the wind speed with a value slightly higher than 14 % and finally, the sunshine duration with a 40 %. The report corroborates that the four variables contribute to the construction of the model; moreover, the type of relationship presented in this figure verifies the behavior assumed in the qualitative analysis where the humidity was described as having an inversely proportional correlation and the rest of variables a directly proportional one.

The last conclusion is reinforced by figure 36 where the relationships temperature - sunshine duration and humidity - wind speed with the output are displayed. When the temperature and the sunshine duration are at the maximum point, the irradiation obtains its maximum value; and when the humidity and the wind speed achieve its minimum and maximum value, respectively, the trajectory of the

sunlight obtains the best conditions for its reception in ground. Therefore, these variables become suitable input elements for the ANFIS estimation model.

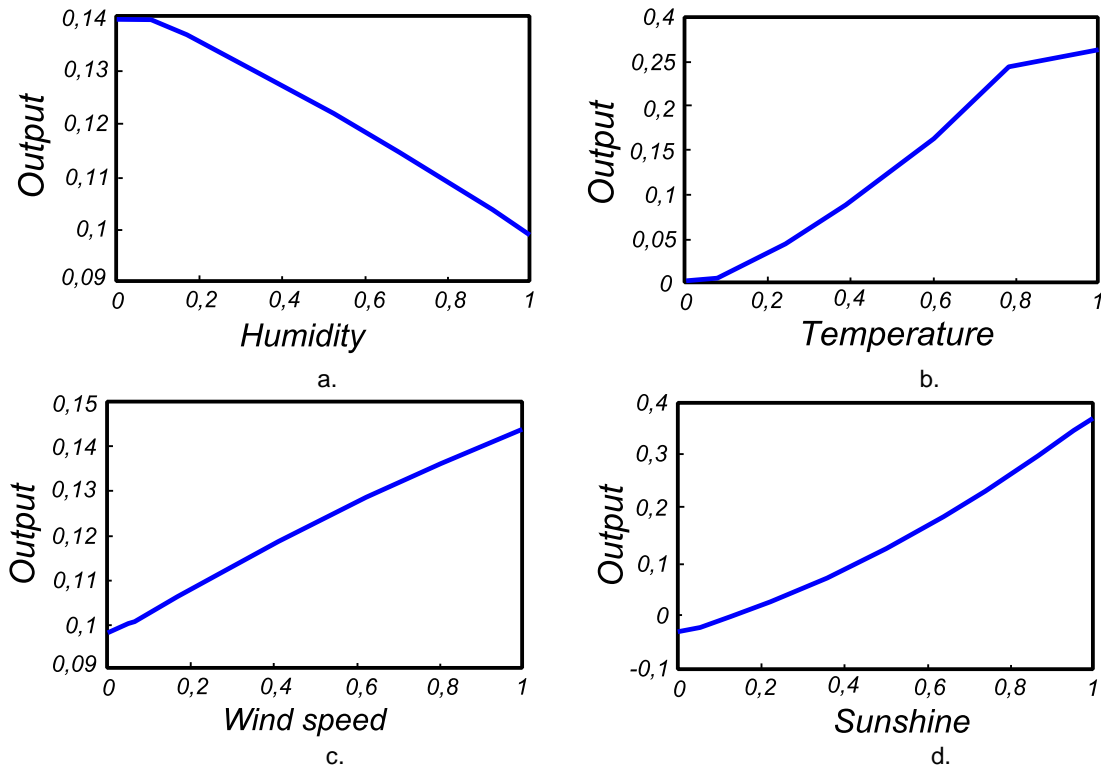


Figure 35. Correlation among input variables with the output in the ANFIS structure.

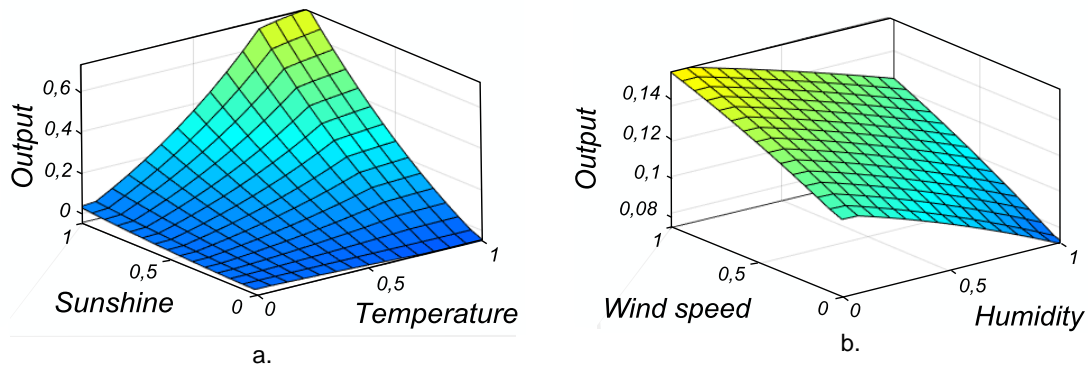


Figure 36. Relationship between input and output data from the qualitative analysis.

- ANFIS error analysis:

As in the ANN analysis, the error indicators were calculated and these are portrayed in table 36, but unlike ANN, this table includes the three time intervals considering that the details and reasons for their implementation were already presented previously. In this sense, table 36 describes again the effect of the estimation range in the accuracy of the model. For higher time scales, the model reduced the distance between the predicted and measured data. Values in MJ/m<sup>2</sup> are not displayed because ANFIS results will not be compared with other works in the literature. That is because the goal of implementing the ANFIS model is to determine if it could contribute with better benefits than the ANN model in the specific solar prediction of the region under assessment.

Since the MAPE indicator is the choice in the present work to compare the performance results, a juxtaposition of this parameter between the ANN and ANFIS models is presented in table 37.

As it is observed in table 37, ANFIS generates slight improvements in the performance of the model in hourly scale; but in the other time series it becomes to reach variations of up 37 % (with a MAPE of 13,9 in ANN model to 8,77 in ANFIS model) in the daily scale for the city of Herrán and results with almost null errors in the monthly scale for the same place. In this way, the ANFIS model could be a more reliable source than the ANN model in several time scales, but its results are not as forceful as for designating to ANFIS above ANN in all cases. It is worth to note that in table 36 and 37, there are not results for the simulation part as in the ANN analysis for monthly scale; this is because the amount of data is limited and was decided to use all data just for the training process.

Although error indicators give a preliminary idea of the performance of the models, the real advantages of these will be analyzed in the next section where their application in the sizing of a PV system is performed.

		Cúcuta		Pamplona		Herrán	
		Training	Simulation	Training	Simulation	Training	Simulation
Hourly	RMSE (Wh/m <sup>2</sup> )	171,9	184,09	105,4	119,59	85,4	107,8
	MAE (Wh/m <sup>2</sup> )	128,3	132,84	69,9	77,9	60,3	70,4
	NRMES	0,162	0,175	0,108	0,113	0,0971	0,1034
	MAPE	18,2	18,7	13,9	142	10,1	11,8
	R <sup>2</sup>	0,889	0,8727	0,898	0,901	0,923	0,9138
Daily	RMSE (Wh/m <sup>2</sup> )	910,4	948,7	794,6	834,38	610,3	729,88
	MAE (Wh/m <sup>2</sup> )	847,6	890,6	680,4	763,4	473,4	555,8
	NRMES	0,158	0,1625	0,112	0,1296	0,0898	0,0955
	MAPE	11,7	12,2	12,01	13,17	8,77	9,3
	R <sup>2</sup>	0,93	0,9265	0,918	0,911	0,933	0,9244
Monthly	RMSE (Wh/m <sup>2</sup> )	47,15	-	96,14	-	0,552	-
	MAE (Wh/m <sup>2</sup> )	27,91	-	58,93	-	0,44	-
	NRMES	0,0011	-	0,0014	-	9,48e-6	-
	MAPE	1,87	-	5,07	-	3,6e-4	-
	R <sup>2</sup>	0,9998	-	0,9998	-	0,9999	-

Table 36. Error indicators for the develop ANFIS models for the evaluated cities in the estimation ranges: Hourly, daily and monthly.

		Cúcuta		Pamplona		Herrán	
		Training	Simulation	Training	Simulation	Training	Simulation
Hourly	ANN	19,55	21,3	17,08	19,7	17,36	18,4
	ANFIS	18,2	18,7	13,9	14,2	10,1	11,8
Daily	ANN	15,39	17,52	15,8	14,29	13,9	12,86
	ANFIS	11,7	12,2	12,01	13,17	8,77	9,3
Monthly	ANN	1,72	-	9,68	-	2,92	-
	ANFIS	1,87	-	5,07	-	3,6e-4	-

Table 37. Comparison between the ANN and ANFIS results in terms of the MAPE indicator.



# Chapter 5

## Comparison for a specific case: PV sizing

After analyzing the error indicator between the developed models and similar works in the literature, it is found that the models have the expected performance, and that ANFIS presents a slight improvement with respect to the model based on ANN. Despite of the above, these results are compared with other irradiation information sources, in order to identify the specific advantages for the sizing of PV systems in the evaluated cities, being this one of the most important goal of the current work.

Considering the information sources in section 4.1.1, the AI models are tested against these sources for the sizing of PV systems, taking as reference the irradiation data provided by IDEAM in hourly scale (called *measured* in tables from here on). The sizing of a grid-connected application for two particular cases (low and high power consumption) and a stand-alone PV system for only one scenario in each city will be used to evaluate the models.

The methodology and results of the comparison are described below.

### 5.1. Grid-connected PV system sizing:

The grid-connected design is based on the energy balance method and the average value of Peak Sun Hours (PSH) included in equation 13.

$$E_{DC} = N_{sG} N_{pG} V_{mMr} I_{mMr} (\overline{PSH}) D \quad Eq. 13$$

Where  $N_{sG}$  is the number of modules in series,  $N_{pG}$  is the number of branches in parallel,  $V_{mMr}$  and  $I_{mMr}$  are the voltage and current coordinates of the Maximum Power Point (MPP) of a single PV module under standard conditions, and  $E_{DC}$  is the DC output energy generated by the PV generator along a period of  $D$  days having an average daily value of peak solar hours defined by equation 14.

$$\overline{PSH} = \frac{\int_{1\text{ year}} G(t) \cdot dt}{(365\text{ days/year})(1000\text{ W/m}^2)} \quad Eq. 14$$

Being  $G(t)$  the global solar radiation on a horizontal surface. Likewise, the DC peak power  $P_{DCpeak}$ , of the PV system can be written as:

$$P_{DCpeak} = N_{sG} N_{pG} V_{mMr} I_{mMr} \quad Eq. 15$$

Which is used to define the efficiency  $\eta$  as follows:

$$\eta = \frac{P_{ACpeak}}{P_{DCpeak}}$$

Being  $P_{ACpeak}$  the AC peak power; this variable must be smaller than the nominal power of the inverter  $P_{nom}$ . Therefore, with the input requirements of the inverter selected for the design (the maximum input current,  $I_{max}$ ) and the PV module to implement (voltage range for a correct MPPT), the number of parallel branches and the amount of PV modules in series for the string, can be calculated as:

$$N_{pG} \leq \frac{I_{max}}{I_{mMr}} \quad Eq. 16$$

$$N_{sG} = \frac{P_{ACpeak}}{\eta N_{pG} V_{mMr} I_{mMr}} \quad Eq. 17$$

Where the value of  $N_{sG}$  must be limited by the range in equation 18.

$$V_{minMPPT} \leq N_{sG} V_{mMr} \leq V_{maxMPPT} \quad Eq. 18$$

The range in the equation 18 avoids: 1) with its low limit, that a part of the  $P_{DC}$  generated by the PV modules will not be converted for the inverter, since when the power at the inverter input is lower than a minimum called *lower threshold voltage*, the inverter output is left open, and 2) with the high limit, to damage physical structure of the inverter due to overload.

The total area for the PV modules is given by equation 19 where  $A_m$  is the individual area of the module.

$$A = N_{sG} N_{pG} A_m \quad Eq. 19$$

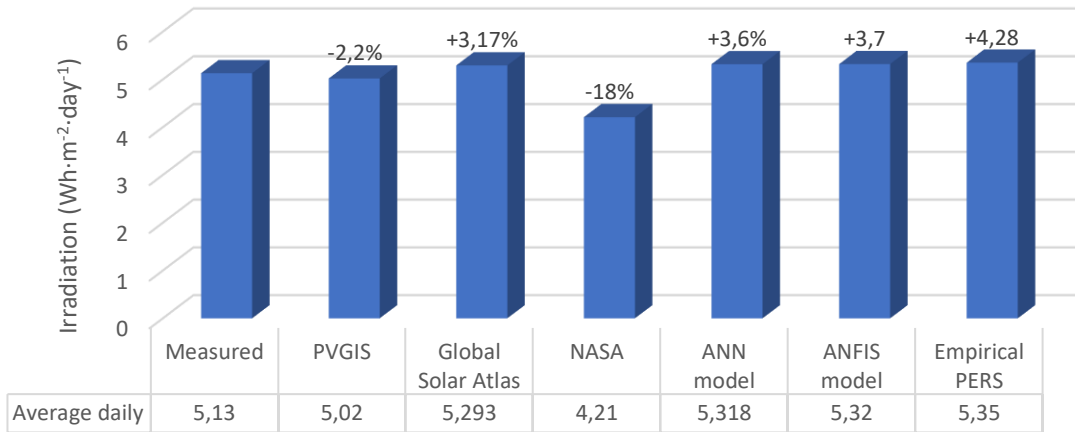
Because the PSH is one of the most important parameters in the presented design method, figure 37 displays it for each one of the analyzed information sources. The PSH used as reference was calculated from the last four years of irradiation data provided by the IDEAM in each city because these years were the most representative by having the largest amount of data per year in a consecutive way. Thus, the data from the other sources were adapted to try of keeping the same conditions respect to the reference and to be able to obtain a reliable comparison.

For Cúcuta, the data from PVGIS, NASA and the AI models have the same conditions for the calculation of PSH; the same days for the years 2012, 2013, 2014 and 2015 were averaged to calculate the PSH values from these sources. PSH for Global Solar Atlas is obtained from its long-term yearly average and for PERS project in a graphical way for the monthly averages values of 2015 presented in the results of its public report. Although these two last sources do not have the same conditions with respect to the other ones, they are included in the analysis in order to identify their impact if these were the only sources a designer would have access to.

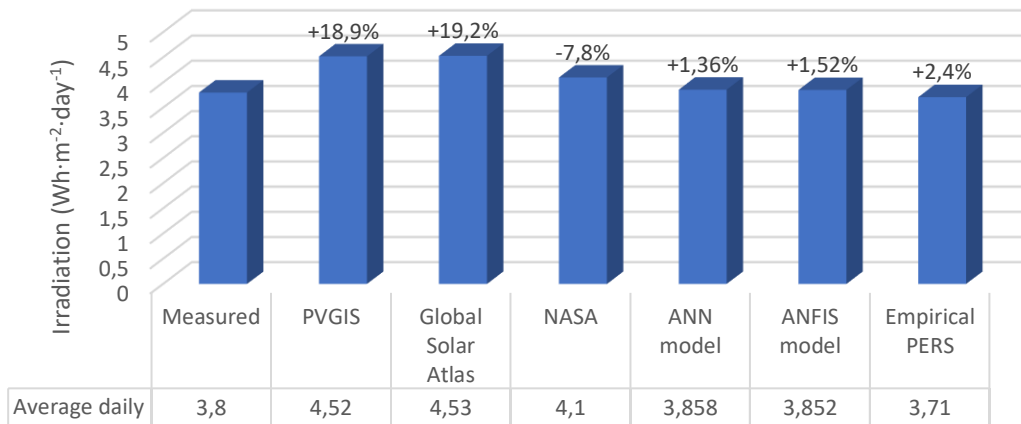
For the cities of Pamplona and Herrán the conditions for the comparison vary more. Since information until 2017 from IDEAM was obtained, the reference value is updated for the last four years until 2017, i.e., reference PSH is the average of the years: 2014, 2015, 2016 and 2017. Similarly, the values for the AI models were also updated. The conditions for the rest to the sources remain equal than for the city of Cúcuta, except for the data provided by the NASA which can also be downloaded for the years 2016 and 2017. Here, it can be remarkable to mention that the reference and AI models value were not calculated for the last years from 2015 to maintain the conditions with PVGIS for the following reasons: 1) To analyze the impact of the updated information assuming that nowadays, a designer only had access to the data from PVGIS; 2) because the data for the years 2012 and 2013 in these cities were not as representative as the years 2016 and 2017 for the calculations.

After the calculations and as it is observed in figure 37, it was found that the NASA supplied the same data for the three cities (graphically it is determined in figure 37b and 37c, since in 37a the data

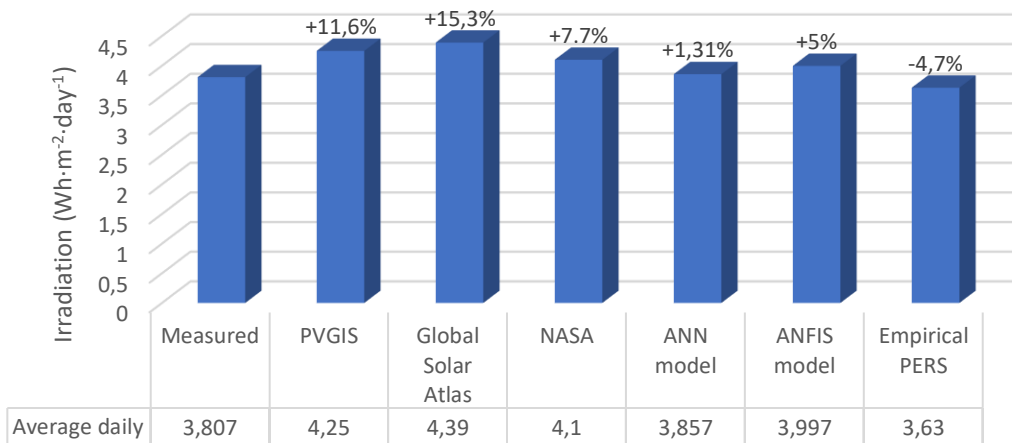
correspond to other years), which does not have a large impact in cities with similar irradiation conditions (as Pamplona and Herrán) but could affect the analysis of the other cities of the region with a much more diverse climate. The percentage of separation respect to the reference also is displayed in figure 37 on top of each bar.



a.



b.



c.

Figure 37. PSH for each one of the information sources in: a. Cúcuta. b. Pamplona. c. Herrán.

To begin the sizing process, two elements must be defined: the PV module and the inverter to use. So, a brief review about the selection process of these is presented in the following sub-sections.

### 5.1.1. PV module selection:

Photovoltaic modules are typically rated between 50 W and 350 W with products specialized for building integrated PV systems (BIPV) at even larger sizes. One of the most used PV modules is the wafer-based crystalline silicon module which have commercial efficiencies between 14 and 22,8 % [98]. Crystalline silicon modules consist of individual PV cells connected together and encapsulated between a transparent front, usually glass, and a backing material, usually plastic or glass.

No specific standards, policy or specifications are available to pick a PV module, but some features can be used as selection guidelines, such as the type of technology (polycrystalline and monocrystalline types being very common), the number of cells (typically between 60 or 72 PV cells connected in series [99]) and the output power according to the application. In the literature for rooftop PV systems there is a wide variety of models, but in [100], where the authors perform a study of the economic feasibility analysis of small scale PV systems in 13 countries, they encourage the use of the SolarWorld 245W Polycrystalline PV module due to its good score in the PV+Test2.0 certification where the durability, electrical safety, workmanship, performance, documentation and guarantee of the PV modules are evaluated. According to [101], this test is a independent certification supported by the German certifier TÜV, with the participation of 21 leading brands, and whose last version presented the results shown in table 38 for the best 9 positions. Additional information about the test can be found in [102].

Assessment	Qualitative result	Manufacturer	Model
92,29	Excellent	Paneles solares SolarWorld	SW260 Poly -> SW290 Mono (Nuevo Standar)
91,3	Excellent (-)	Paneles solares Schott	Schott Poly 290
90,7	Excellent (-)	SHARP solar	NU-180E1
89,8	Good (+)	IBC	IBC Monosol 240 ET
89,0	Good (+)	Mitsubishi Electric	PV-TD185MF5
88,5	Good (+)	Jetion	JT235PCe
88,1	Good (+)	Conergy	PowerPlus 225P
84,3	Good	Sovello	SV-X-195-fa1
80,0	Good (-)	Perfect Solar	PS230-6P-TOP

Table 38. Results of the PV+Test for the top 9 PV panels.

The PV + Test score grades range from the highest, which is Excellent, to the Very Poor. In each grade, depending on the results, you can carry (+) or (-) as a classification of intermediate grades.

Under this criterion, the PV module type SolarWorld SW260 Poly was the choice for the analysis; its main operating characteristics under Standard Test Conditions (STC) are presented in table 39.

### 5.1.2. Selection of the inverter:

The inverter should fulfill the waveform requirements of the load or grid, and have a high efficiency in order to reduce the loss of energy generated by the PV modules. In [100], [99], [103], [104] and [105], the authors coincide in that the Sunny production line from SMA Solar Technology AG is an adequate option for residential applications, and even for commercial systems; hence, the Sunny

Boy 3000TL-US inverter is selected and its characteristics are depicted in table 40. A wide operating range, efficiency close 100 % and enough nominal power for the system are the principal selection criteria for this device. A more detailed analysis for the inverter selection must be performed for a real implementation but this is not the goal of this section, since the main objective is to identify the variation in the sizing of a standard system by using different solar radiation sources.

Parameter	Value
Output power	290 Wp
MPP Voltage	31,9 V
MPP Current	9,2 A
Open circuit voltage	39,6 V
Short-circuit current	9,75 A
Efficiency	17,3 %
Number of cells	60
Length	1675 mm
Width	1001 mm
Height	33 mm
Output tolerance	+/-2,0 %

Table 39. Characteristics of the selected PV module.

Variable	Description
Nominal output AC power (kW)	3
Maximum AC Apparent Power (kVA)	3
Rated MPPT voltage range (V)	175 – 480
Maximum input DC voltage (V)	600
Maximum input DC current (A)	15
Lower threshold voltage (V)	125
AC grid frequency (Hz)	50/60
MPP Trackers	2
Maximum efficiency (%)	97.2
Harmonics (%)	<4

Table 40. Technical data for the selected inverter.

Therefore, considering the information in figure 37 and the characteristics of the elements of the system (PV module and inverter), table 41 shows the different PV sizes in terms of the number of modules for each one of the information sources, taking into account the previous expressions (Eq. 13 to Eq. 19). The load used in table 41 of 8 kW and 24 kW correspond to the daily average power consumption of two residences in the city of Cúcuta; the first scenario, considered a low consumption, and the second one a high consumption; since the latter contains several equipment not common in a typical house of the city, such as air conditioners, electrical garage, dryer, among others, which increase the normal power requirements of a house in Cúcuta.

As observed in table 41a, the sizing for the city of Cúcuta does not have a large impact if the information is taken from a source or another, except for the NASA database, with an oversizing of about 20 % as it was expected from the calculation of its PSH. Despite of the above, for the other cities, the NASA obtains a correct sizing with respect to the irradiation values measured from IDEAM, and the variations are obtained for the data from PVGIS, GSA and empirical models generated by

PERS, with an oversizing of close 15 % for the first two sources (1 and 3 modules, for the 8 kW and 24 kW system, respectively) and an undersizing of the same value for the last one.

	PSH	8 kW system			24 kW system		
		Modules in series	Parallel branches	Area (m <sup>2</sup> )	Modules in series	Parallel branches	Area (m <sup>2</sup> )
Measured	5,13	5	1	8,38	5	3	25,15
PVGIS	5,02	5		8,38	5		25,15
GSA	5,293	5		8,38	5		25,15
NASA	4,21	6		10,06	6		30,18
ANN model	5,318	5		8,38	5		25,15
ANFIS model	5,32	5		8,38	5		25,15
Empirical model	5,35	5		8,38	5		25,15

a.

	PSH	8 kW system			24 kW system		
		Modules in series	Parallel branches	Area (m <sup>2</sup> )	Modules in series	Parallel branches	Area (m <sup>2</sup> )
Measured	3,8	7	1	11,74	7	3	35,21
PVGIS	4,52	6		10,06	6		30,18
GSA	4,53	6		10,06	6		30,18
NASA	4,1	7		11,74	7		35,21
ANN model	3,858	7		11,74	7		35,21
ANFIS model	3,852	7		11,74	7		35,21
Empirical model	3,71	7		11,74	7		35,21

b.

	PSH	8 kW system			24 kW system		
		Modules in series	Parallel branches	Area (m <sup>2</sup> )	Modules in series	Parallel branches	Area (m <sup>2</sup> )
Measured	3,807	7	1	11,74	7	3	35,21
PVGIS	4,25	6		10,06	6		30,18
GSA	4,39	6		10,06	6		30,18
NASA	4,1	7		11,74	7		35,21
ANN model	3,857	7		11,74	7		35,21
ANFIS model	3,997	7		11,74	7		35,21
Empirical model	3,63	8		13,41	8		40,24

c.

Table 41. PV system sizing for a residential application for the city of: a. Cúcuta. b. Pamplona. c. Herrán.

In this way, the first conclusion from table 41 is that the AI models are suitable for proper sizing in all the scenarios under analysis. Secondly, the participation of the updated data has an impact over the sizing which represents an advantage of the AI models with respect to sources as PVGIS, that for an assessment in the same period shows very good results (case of Cúcuta) but for a period out of its range not so much.

In spite of the previous analysis, it is known that the accuracy in the sizing of a grid-connected PV system is not an indispensable factor for the requirements of its application because the grid acts as support in the case of that the energy from the PV system fails by low irradiation (or other external factor). Thus, in this type of system, the correct sizing plays a more economic role applied to the exchange of energy with the grid, that although is also important, is not fundamental for the operation of the load or even for the confort of the final user. In this way, a high precision sizing is more necessary in off-grid systems, and for this reason, this scenario is analyzed in more detail below.

## 5.2. Stand-alone PV system sizing:

Autonomous or stand-alone photovoltaic systems (SAPVS) are installations with photovoltaic modules and batteries designed to meet some load without any connection to the electric grid [106]. In the literature, various methods for the sizing of this type of system have been developed, which differ in terms of their simplicity or reliability. The construction of the sizing curve based on the loss of load probability (LLP), is one of these; it is characterized by its high reliability and accuracy in the design, but also by the need of using solar radiation measurements in the long term which are not always available. Since the first features are fundamental in the present research study, this is the method implemented for the sizing of the stand-alone system and its step-by-step method is explained below.

The merit of a SAPVS should be judged in terms of the realibility of the electricity supply to the load. This is usually quantified by the concept of *loss of load probability*, defined as the ratio between the energy deficit and the energy demands both on the load, over a long period of time [107]. In statistical terms, the LLP value refers to the probability of the system to be unable to meet the demand. The main reason for this failure is the stochastic characteristics of the solar radiation that affect the sizing process [108].

A LLP analysis for SAPVS sizing is characterized by two dimensionless parameters according to [107]:  $C_S$ , related to the capacity of the storage system, and  $C_A$ , the mean or minimum capacity of the PV panels array, defined as:

$$C_S = \frac{C_U}{L} = \frac{N_B V_B C_B DOD}{L} \quad Eq. 20$$

$$C_A = \frac{\eta A \overline{G_d(\beta)}}{L} \quad Eq. 21$$

where  $N_B$  is the number of batteries (supposed all equal),  $V_B$  the nominal voltage of one battery (in V),  $C_B$  the nominal or rated capacity of each battery (in  $A \cdot s$ ),  $DOD$  the maximum allowable depth of discharge of each battery (dimensionless),  $C_U$  the maximum useful capacity of the batteries (in W),  $L$  the mean daily energy load (in W),  $\eta$  is the average whole energy transmission efficiency of the PV system from the PV array to the load,  $A$  the array area (in  $m^2$ ), and  $\overline{G_d(\beta)}$  is a representative daily insolation on the plane (with tilt  $\beta$ ) of the array (in  $W/m^2$ ).

The physical sense of  $C_S$  and  $C_A$  is clear:  $C_S$  represents the number of days the batteries are at full capacity and with no energy income could feed the load (assumed constant), while  $C_A$  means the number of loads (each one with value L) that are expected to be fed by the PV array alone. This work is concerned with fixed tilt plants (no Sun tracking), so  $\beta$  is supposed to be constant for a certain installation [106].

Given a location and a load, two general ideas are intuitive: Firstly, it is possible to find many different combinations of  $C_A$  and  $C_S$  leading to the same Loss of load probability (LLP) value. Secondly, the larger the PV-system is, the greater the cost and the lower the LLP [109]. The traditional problem given by Egido and Lorenzo [107] is formulated as follows: *Which pair of  $C_A$  and  $C_S$  values lead to a given LLP value at the minimum cost?*

In order to give solution to the previous question, several methods are shown in the literature in how to calculate the optimal sizing parameters for a constant LLP by means of the graphical representation of  $C_A$  respect to  $C_S$ . Each point of the  $C_A - C_S$  plane represents a size of a PV system. This allows to obtain the reliability map as figure 38 shows. The lines are the loci of all the points corresponding to the same LLP value. Because of that, these are called *isoreliability lines*. The definition of  $C_A$  and  $C_S$  implies that this map is independent of the load and depends only on the meteorological behavior of the location. It will be observed that the isoreliability lines are, very close to, a hyperbola with their asymptotes parallel to the  $x$  and  $y$  axis, respectively.

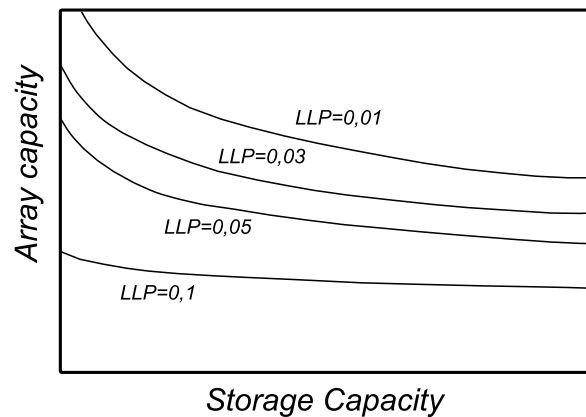


Figure 38. Reliability map.

On the other hand, given an LLP value, the plot of the cost of the PV system corresponding to the isoreliability line is, approximately, a parabola having a minimum that defines the *optimal solution* to the sizing problem [107], as it is observed in figure 39.

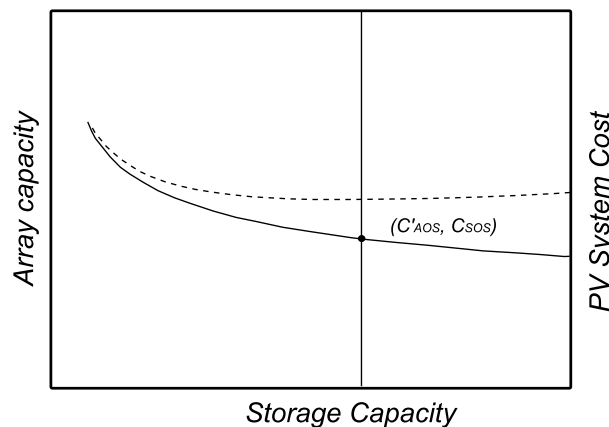


Figure 39. Iso-reliability curve for a specific LLP with its corresponding plot of cost indicating the optimal pair  $C_{AoS}, C_{SOS}$ .

The methods to apply the LLP concept include: *intuitive methods*, where the size of the system is taken in such way to ensure the load demand without giving a relation between the number of panels, batteries and the LLP. *Analytical methods*, based on the graphical information obtained from the iso-



probability curves, and the *numerical methods*, which are based on a detailed simulation of the PV system in small scales (daily, hourly, etc.). The advantages of the numerical methods are the precision and the simplicity of choosing different elements of the system [108], so that is the choice to calculate the reliability map in this work.

### 5.2.1. Numerical method:

As it was indicated, the pair of  $C_A$  and  $C_S$  in this method is calculated by means of a rather detailed simulation of the PV system. To explain this, the system showed in figure 40 is analyzed; it is assumed that all the consumption occurs during night and that the battery is free of energy losses.

Based on the above, the auxiliar generator operates as a support to the battery and to the load, when the energy provided by the PV array is not enough. If the simulation is carried out over a great number of days  $N$ , in order to be statistically meaningful, then the LLP value for the stand-alone system (i.e. the system of the figure 40, excluding the auxiliar generator) is given by equation 22:

$$LLP = \frac{\sum_{j=1}^N E_{AUXj}}{\sum_{j=1}^N L_j} \quad Eq. 22$$

Where  $j$  indicates the day in evaluation.

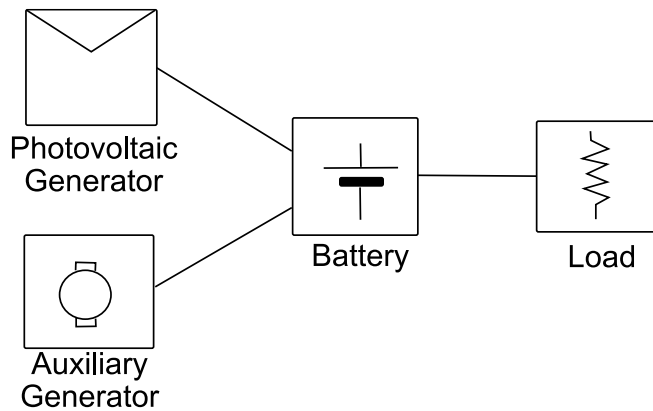


Figure 40. Schematic of the photovoltaic system used for the numerical method.

Because long term averages of daily irradiation incident on surfaces other than horizontal are not generally available, it is useful to define a parameter to relate it:

$$C'_A = \frac{\eta A \overline{G_d(0)}}{L} = C_A \frac{\overline{G_d(0)}}{\overline{G_d(\beta)}}$$

Where  $\overline{G_d(0)}$  is the mean daily irradiation on an horizontal surface. In practical cases, PV designers can directly obtain the value of  $\eta A$  (and consequently, the required number of PV modules) from  $C'_A$ .

In general, since the long term irradiation data are common in a monthly or annual basis, the daily irradiation data for the expressions above are the result of averages in these periods; moreover, in [107], the authors indicated that the location dependence of the results can be reduced by correlating the collector area to the irradiation *worst month*, defined as the month with the lowest relation between solar irradiation and energy consumption.

The simulation algorithm to identify the behavior of the system is structured in the analysis of the *state of charge (SOC)* in the battery, and it has different derivations from the one shown by Egido

and Lorenzo [107] in the literature [109] [106] [110] [111]. In the present work, the algorithm developed for the calculation of the pairs  $C_A, C_S$  is shown below. It is valid to use a constant load to obtain a first perspective of the system behavior; in addition to this, the following analysis implies the use of a constant PV array efficiency, which is supported in a research study referenced in [107], where they indicated that for the same LLP value using daily timesteps and a constant efficiency, similar results were obtained than using hourly timesteps and considering the efficiency dependence on solar irradiance, ambiente temperature and battery state of charge as shown in equation 23, particularly for  $C_S \geq 2$ :

$$\eta_t = \eta_p(1 - B \cdot (T_c - T_r)) \quad Eq. 23$$

Where,  $\eta_p$  is the efficiency of a solar cell at a referenced solar radiation (i.e. 1000 W/m<sup>2</sup>),  $B$  is the temperature coefficient (between 0,004 and 0,006),  $T_c$  is the cell temperature in °C and  $T_r$  is the reference temperature of the panel (generally equal to 25 °C with air mass AM = 1,5).

Despite of this, a comparison between these timesteps is evaluated in next sections. The steps of the curves generation algorithm are the following:

*Step 1:* Initialization and loading the data

$$LLPs = 0,01, \quad L = 8kW, \quad \eta = 0,175, \quad G = \text{Daily irradiation data}, \quad C_S = 1, \quad C_A = 1, \\ SOC_1 = 0, \quad E_{AUX_1} = 0, \quad Err = 1 \times 10^{-3}$$

*Step 2:* Calculate the fixed variable in the time period

$$A = \frac{C_A \bar{L}}{\eta \bar{G}_d}, \quad C_U = C_S \bar{L}$$

*Step 3:* Calculate the state of charge of the battery for each day

$$SOC_d = \min\left(SOC_{d-1} + \frac{\eta A G_d}{C_U}; 1\right)$$

The state of charge is obtained from its value for the previous day plus the energy produced by the array during the day. After this, the energy supplied by the auxiliar generator keeps the battery charged to the level of the load if the stored energy is lower than its requirements:

$$SOC_d \geq \frac{1}{C_S} \Rightarrow E_{AUX_d} = 0 \\ SOC_d < \frac{1}{C_S} \Rightarrow E_{AUX_d} = \left(\frac{1}{C_S} - SOC_d\right) * L_d * C_S \quad \text{and} \quad SOC_d = \frac{1}{C_S}$$

Finally, the state of charge of the battery at the end of the day is established:

$$SOC_d = SOC_d - \frac{L_d}{C_U}$$

*Step 4:* Calculate the LLP value from equation 22 for the total data.

*Step 5:* Determine the error with respect to the desired LLPs to continue or stop the iteration process. If  $abs(LLP - LLPs) > Err$ , then go to the step 6; else go to the step 7.

*Step 6:* Define if the current LLP value is above or below of the desired LLPs to apply the increase or decrease in the  $C_A$  value. *If  $LLP > LLPs$ , then  $C_A = C_A + 0.001$ ; else  $C_A = C_A - 0.001$ .* Go to step 2.

*Step 7:* Save the pair of  $C_A, C_S$ , and increase the  $C_S$  value. *If  $C_S = 10$ , then stop the algorithm. else go to the step 2.*

The previous algorithm follows the original methodology presented by Egido and Lorenzo in [107], but assuming the strategy applied by Lucio et. al. in [106] in which the auxiliar generator supplies only the energy left to  $L$  (no more to fully charge the battery), and the operation of the algorithm used by Mellit in [109].

The simulations were performed for typical values of LLP of 0,1, 0,5 and 0,01 and the results for each city are shown in figure 41. The data used to construct the curves correspond to the period between 2012 and 2015 for Cúcuta, and 2014 - 2017 for Pamplona and Herrán. In general, the worst month in terms of irradiation is selected to apply the method, which is usually December in the Northern Hemisphere; but due to the proximity to the Equator line of the places under evaluation, or otherwise, because the lack of the seasons, is not very clear to select a specific month as that with the lowest irradiation year after year. For instance, for the city of Cúcuta, March in 2012 was one of the months with the lowest irradiations of that year, but in 2014 it was one of the months with the highest irradiation. While in 2013 and 2015 irradiation values for March were located in the average of those years. In addition to this, the variation from month to month do not change as much as in countries with seasons in the Northern and Southern Hemispheres.

Therefore, the months with the lowest irradiation among the years analyzed for each city were evaluated; in some cases, two or three months in the year showed the lowest irradiation with small differences, so we determined a limit to categorize a month with this characteristic as it is indicated in table 42. Thus, from this information the days with low irradiation can also be categorized considering the monthly average daily irradiation in each case (showed in table 42 too).

<b>Period of classification for low irradiation</b>	<b>Cúcuta (kWh/m<sup>2</sup>)</b>	<b>Pamplona (kWh/m<sup>2</sup>)</b>	<b>Herrán (kWh/m<sup>2</sup>)</b>
Monthly limit	<138	<100	<100
Daily limit	<4,6	<3,4	<3,4

Table 42. Limits to classify a month and a day with low irradiation.

In this way, in order to cover the behavior of the whole year not only were selected the days inside of the months with the lowest irradiation, but also the days during the year with values lower than the limit established in table 42. Therefore, bad weather conditions in other periods of the year which affect the solar radiation are analyzed to increase the accuracy of the curve and the final sizing of the system.

Considering the concept given by Egido and Lorenzo for the optimal pair of  $C_A, C_S$ , table 43 shows just an approximation of these values for each city, since many combinations of PV modules and batteries can be evaluated for optimizing, in economic terms, the values of  $C_A$  and  $C_S$  (as in [108]), which exceeds the scope of the current study. Furthermore, figure 41 depicts the LLP for the same cities. To identify the accuracy of the information sources, their isoreliability curves for a LLP of 0,01, 0,05 and 0,1 for each city are portrayed in figure 42, and several sizings based on these curves are described in table 44. The battery Trojan J305G was used for the calculations, with a nominal capacity of 315 Ah and a voltage of 12 V.

From table 44, we can observe that the sizing based on the lowest irradiation days presents small variation among the different information sources and the reference. Data from PVGIS showed the largest amount of differences (highlighted in light red in table 44) for the sizing in the three cities although for the city of Cúcuta the evaluation period was the same that the one used by the rest of the sources (as it has been indicated during this section). Data from NASA on the other hand, using the same value for the city of Pamplona and Herrán (which could have been the same for Cúcuta if the evaluation period was also the same), obtained good results for the three cities.

A first comparison between the LLP sizing method and the energy balance method used in the grid-connected application allows us to conclude that data from NASA has a higher precision when it uses the days with the worst irradiation conditions and data from PVGIS when uses sizing methods with average annual values. In the same way, AI models obtained good results in both sizing processes with the least variations with respect to the reference.

LLP	Cúcuta		Pamplona		Herrán	
	$C_A$	$C_S$	$C_A$	$C_S$	$C_A$	$C_S$
0,01	1,08	2,3	1,04	2	1,06	2
0,05	0,97	1,8	0,97	1,8	0,967	1,8
0,1	0,9	1,46	0,9	1,48	0,9	1,6

Table 43. Optimal pairs  $C_A, C_S$  for different LLP values in daily scale analysis.

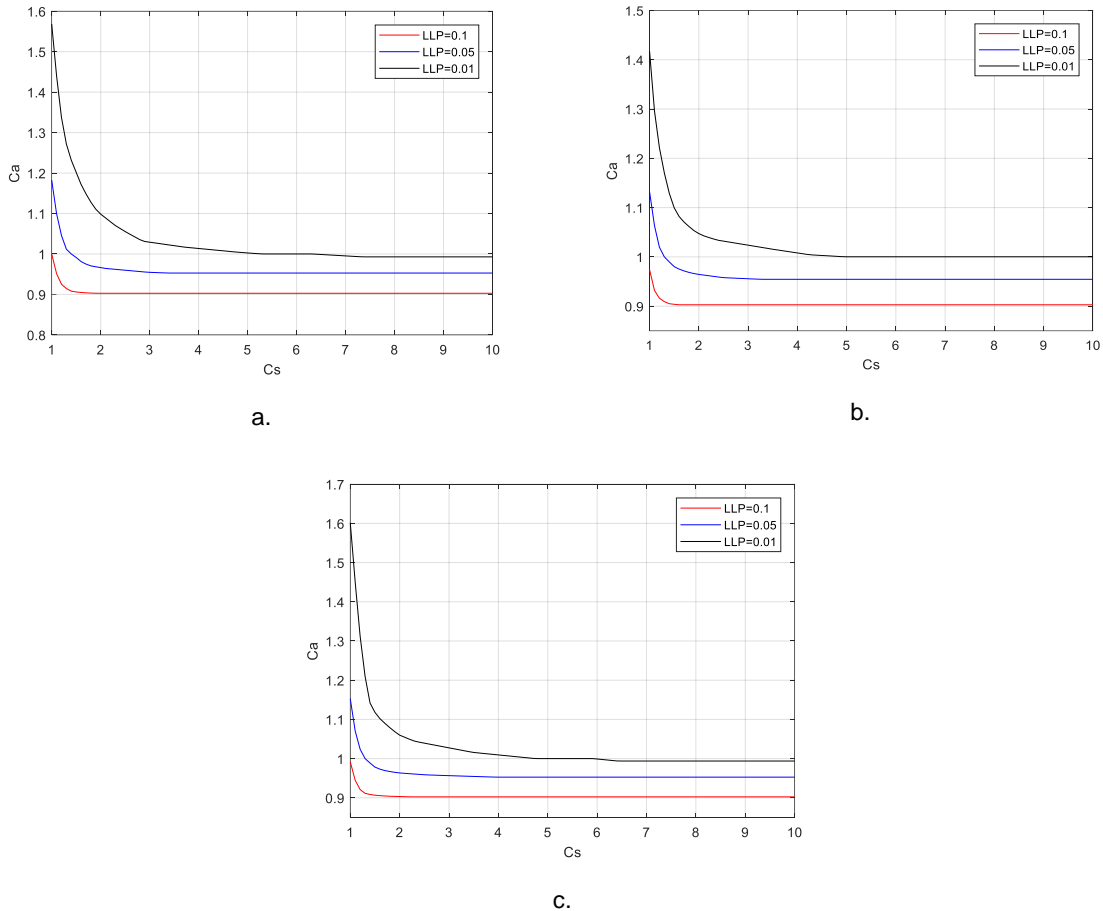
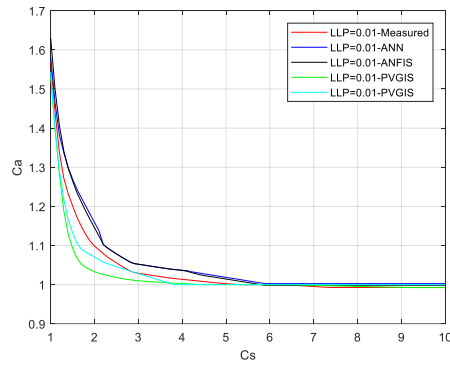
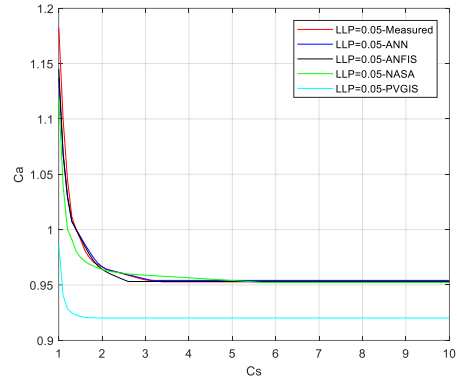
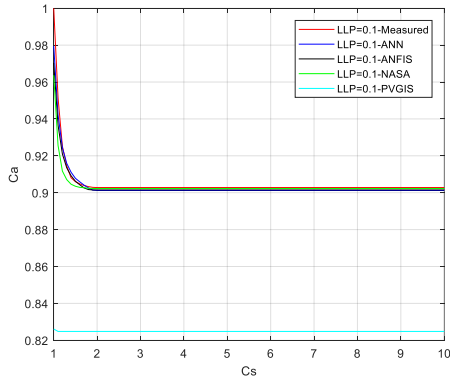
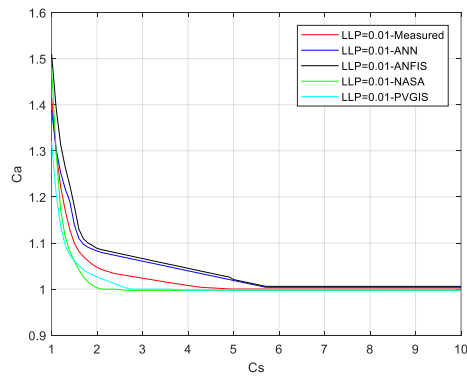
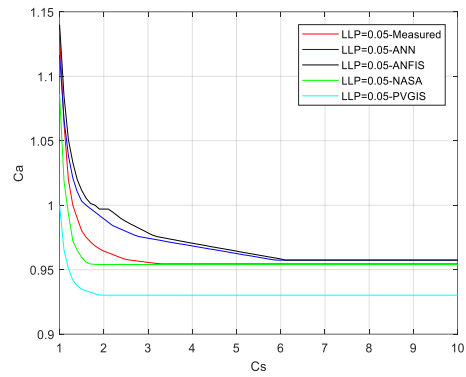
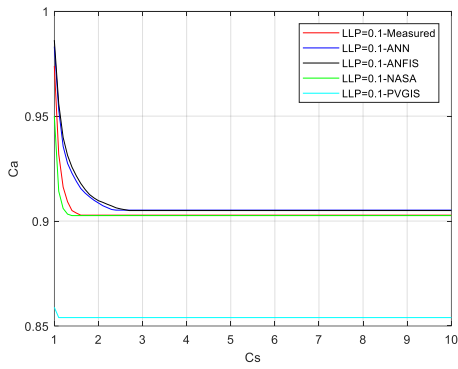


Figure 41. Iso-reliability curves for the cities: a. Cúcuta. b. Pamplona. c. Herrán.



a.



b.

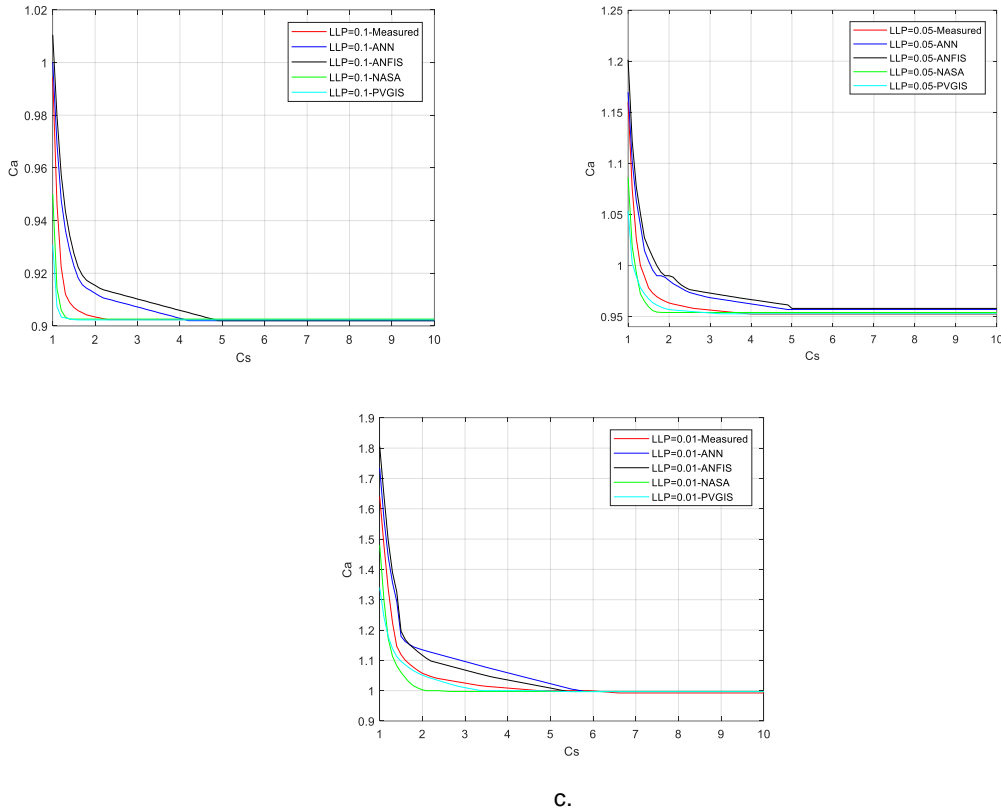


Figure 42. Iso-reliability curves comparison for the different information sources: a. Cúcuta. b. Pamplona. c. Herrán.

	Cs	Measured		PVGIS	NASA	ANN	ANFIS
		Number of panels	Number of batteries	Number of panels	Number of panels	Number of panels	Number of panels
0,01	1	13	2	12	12	13	13
	2	9	4	9	8	10	9
	3	8	6	8	8	8	8
	5	8	11	8	8	8	8
	10	8	21	8	8	8	8
0,5	1	10	2	8	9	9	9
	2	8	4	7	8	8	8
	3	8	6	7	8	8	8
	5	8	11	7	8	8	8
	10	8	21	7	8	8	8
0,1	1	8	2	7	8	8	8
	2	7	4	7	7	7	7
	3	7	6	7	7	7	7
	5	7	11	7	7	7	7
	10	7	21	7	7	7	7

a.

	Cs	Measured		PVGIS	NASA	ANN	ANFIS
		Number of panels	Number of batteries	Number of panels	Number of panels	Number of panels	Number of panels
0,01	1	15	2	14	16	15	16
	2	11	4	11	11	11	11
	3	11	6	11	11	11	11
	5	11	11	11	11	11	11
	10	11	21	11	11	11	11
0,5	1	12	2	11	12	12	12
	2	10	4	10	10	10	10
	3	10	6	10	10	10	10
	5	10	11	10	10	10	10
	10	10	21	10	10	10	10
0,1	1	10	2	9	10	11	11
	2	10	4	9	10	10	10
	3	10	6	9	10	10	10
	5	10	11	9	10	10	10
	10	10	21	9	10	10	10

b.

	Cs	Measured		PVGIS	NASA	ANN	ANFIS
		Number of panels	Number of batteries	Number of panels	Number of panels	Number of panels	Number of panels
0,01	1	18	2	15	16	19	20
	2	11	4	11	11	12	12
	3	11	6	11	11	12	11
	5	11	11	11	11	11	11
	10	11	21	11	11	11	11
0,5	1	13	2	11	12	13	13
	2	10	4	10	10	10	11
	3	10	6	10	10	10	10
	5	10	11	10	10	10	10
	10	10	21	10	10	10	10
0,1	1	11	2	10	10	11	11
	2	10	4	10	10	10	10
	3	10	6	10	10	10	10
	5	10	11	10	10	10	10
	10	10	21	10	10	10	10

c.

Table 44. PV system sizes for the different information sources in terms of the autonomy days for a constant load of 8 kW: a. Cúcuta. b. Pamplona. c. Herrán.

Now, although Egido and Lorenzo found in a previous research study that the hourly timestep compared the daily one has similar results, Benmouiza et. al. in [108] showed that the sizing based on an hourly analysis can generate better results than the daily or monthly scales, indicating that

daily analysis is not always the best choice because it does not give all the dynamic characteristics of the solar radiation and does not contemplate the presence of bad weather conditions, which can occur even outside the worst month of the year. Thus, following the methodology in [108], hourly isoreliability curves were obtained for each city to determine if that improves the sizing results, which would represent an advantage of the ANN and ANFIS models with respect to the other information sources, since these models can provide hourly information that most of the other sources can not (except the PVGIS source which also supplies information in hourly scale, but even so, it is outdated respect to the artificial intelligence models).

The methodology is the following: First, time series data mining was applied to hourly solar radiation data. It consists of grouping similar elements into clusters that have the same characteristics in order to obtain the hours with the lowest irradiation inside the total data provided by the IDEAM, being this the equivalent step to the selection of the worst irradiation month used in the daily scale analysis. For this purpose, several methods have been proposed in the literature based on unsupervised clustering methods such as k-means and fuzzy c-means (FCM) algorithms; the latter was the one selected because fuzzy c-means presents more precise results compared with k-means algorithm [108].

Due to the nonlinearity of the solar radiation data, before applying the FCM method, *phase space reconstruction* is implemented to overcome this issue. Phase space reconstruction is a technique which presents the data in high dimensional space based on Takens theorem for a better understanding and analysis of the underlying dynamical of the system. It consists of determining the minimum embedding dimension for a time series, being the time delay embedding method one of its most common versions [108]. A scalar time series  $x(t_i)$  is embedded into an  $m$ -dimensional space denoted  $X(t_i)$ ; as it is expressed in equation 24:

$$X(t_i) = x(t_i), x(t_i + \tau), \dots, x(t_i + (m - 1)\tau) \quad Eq. 24$$

where,  $i = (1, 2, \dots, M)$ ,  $\tau$  is the delay time,  $m$  is the embedding dimension and  $M$  is the number of embedded points in the  $m$ -dimensional space given by equation 25.

$$M = N - (m - 1)\tau \quad Eq. 25$$

where,  $N$  is the total number of points of the time series and  $X(t_i)$  is the embedded time series into an  $m$ -dimensional space. As it is observed, to use the concept of the phase space reconstruction, two parameters must be defined: the number of delay and the embedding dimensions ( $m$ ). For the first, the mutual information method proposed by Fraser and Swinney was used [108]. The optimum delay is equal to the first minimum of the plotted mutual information expressed by the following equation:

$$I(x(t), x(t - \tau)) = \sum_{x \in \chi} \sum_{y \in \gamma} p(x(t), x(t - \tau)) \log \frac{p(x(t), x(t - \tau))}{p(x(t))p((t - \tau))}$$

Where  $I(x(t), x(t - \tau))$  is the mutual information and  $p(x(t), x(t - \tau))$  is the joint probability mass function for the marginal probability mass functions  $x(t)$  and  $x(t - \tau)$ . For the second one, the false nearest neighbour method was used to choose the suitable number of the embedding dimension [108]. This method determines the nearest neighbour of every point in a given dimension, and then checks if there are close neighbors in the higher dimension.

Both methods were executed in Matlab from algorithms available in its repository; hourly data of the last two years provided by the IDEAM for each city were used, despite of in [108] only used one year of data; these are: 2014-2015 for Cúcuta, and 2016-2017 for Pamplona and Herrán. Unlike the daily



scale analysis, the hourly method does not require a long term of data since it analyzes the behavior during all days covering different situations in the year. In a similar way to [108], it was found that a delay of one and two embedding dimensions are enough for the phase space reconstruction of the time series evaluated as it is shown in figure 43.

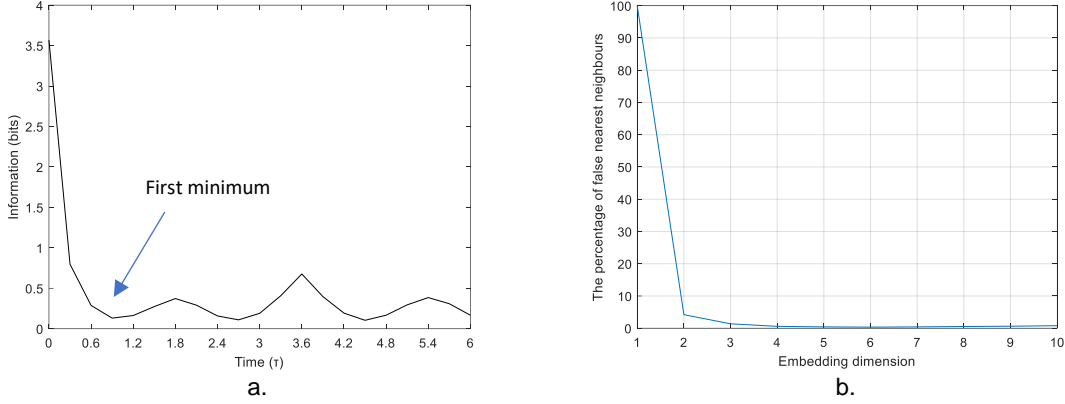


Figure 43. Parameters for the phase space reconstruction: a. Number of delay. b. Minimum embedding dimensions.

With the parameters defined for the set of data, the reconstructed phase space of the solar radiation data is clustered using the fuzzy c-means algorithm. In this method, as in the fuzzy logic approach, each point belongs to a cluster with some degree of belonging defined by a membership grade. The FCM algorithm minimizes an objective function  $J_{FCM}$  that calculated the weighted within-group sum of squared errors as expressed in equation 26:

$$J_{FCM} = \sum_{k=1}^n \sum_{i=1}^c u_{ik}^q \cdot d^2(x_k, v_i) \quad \text{Eq.26}$$

where,  $n$  is the length of the data,  $c$  is the number of clusters defined by the c-means algorithm,  $u_{ik}$  is the degree of membership of  $x_k$  in the  $i_{th}$  cluster,  $q$  is a weighting exponent on each fuzzy membership, it is a real number greater than 1 (typically  $q = 2$ ),  $X = (x_1, x_2, \dots, x_n)$  is the data in the  $m$ -dimensional vector space,  $v_i$  is the center of the cluster  $i$ ,  $d^2(x_k, v_i)$  is the distance measured between data  $x_k$  and cluster center  $v_i$ .

The summary of the FCM algorithm is illustrated by the following steps:

1. Initialize the values  $c, q$  and the error  $\varepsilon$ .
2. Initialize the cluster centre matrix  $V^{(t=0)} = [v_i^{(t=0)}]$  and the membership matrix  $U^{(t=0)} = [u_{ik}^{(t=0)}]$ .
3. Increase the time  $t$  and calculate the new  $c$  cluster centers  $V^t$ :

$$V^t = \frac{\sum_{k=1}^n ((u_{ik})^{(t)})^q x_k}{\sum_{k=1}^n ((u_{ik})^{(t)})^q}$$

4. Calculate the new membership values  $U^{(t+1)}$ :

$$U^{(t+1)} = [u_{ik}^{(t=0)}] \frac{1}{\sum_{j=1}^c \left( \frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{q-1}}}$$

Where  $d_{ik} = \|x_k - v_i\|$ , and  $1 \leq k \leq n$ ;  $1 \leq i \leq c$ .

5. If  $\|U^{(t)} - U^{(t+1)}\| < \varepsilon$  stop. Otherwise, increase  $t$  and go to Step (3).

The FCM algorithm depends strongly on the position of the initialization points. Hence, an important task in the FCM algorithm is choosing the correct number of clusters in order to avoid the problem of falling in a local minimum. Several techniques were proposed in the literature to solve this problem, such as the sub-clustering method and the mountain method. As in [108], the sub-clustering technique was used to decide the number of clusters. This method is an iterative process that supposes each point is a potential cluster center according to its location to other data points. The algorithm is summarized as follows:

- Choose a point that has the probability of being the highest potential cluster center.
- Delete all the points which are inside the radius of the first cluster center (the radius is defined by the neighborhoods of the center), and recalculate the potential of the other points to determine the next cluster center.
- Repeat this step until all the data are within the radius of a cluster center.

Matlab provides a graphic GUI to apply the sub-clustering method and later, the FCM algorithm. For the first method, it is required a *.dat* file with the data to be clustered, specified as an  $M - by - N$  array, where  $M$  is the number of data points and  $N$  is the number of data dimensions with their delays. Thus, the number of rows of the array is given by  $Rows = total\_data - (m \times \tau)$ . To obtain the centers, other parameters are included by Matlab:

- Influence range: Range of influence of the cluster center for each input and output assuming the data falls within a unit hyperbox, with values between 0 and 1. Specifying a smaller range of influence usually creates more and smaller data clusters, producing more fuzzy rules.
- Squash factor: It is used for scaling the range of influence of cluster centers, specified as a positive scalar. A smaller squash factor reduces the potential for outlying points to be considered as part of a cluster, which usually creates more and smaller data clusters.
- Acceptance ratio: Defined as a fraction of the potential of the first cluster center, above which another data point is accepted as a cluster center, specified as a scalar value in the range [0, 1]. The acceptance ratio must be greater than the rejection ratio.
- Rejection ratio: Defined as a fraction of the potential of the first cluster center, below which another data point is rejected as a cluster center, specified as a scalar value in the range [0, 1]. The rejection ratio must be less than the acceptance ratio.

In the order as they appear above, the value for each parameter implemented in the simulation were: 0,7, 1,25, 0,5 and 0,15. For each city, the number of centers obtained was 3 by dimension, which was used to initialize the FCM algorithm. The results from the FCM algorithm in each city are shown in figure 44, where the phase space reconstruction for the hourly irradiation data at time  $t$  and  $t + 1$  are represented. As it is observed in figure 49, the clustering has been performed in three groups which clearly define three irradiation levels: low, medium and high solar radiation with a corresponding hourly pattern. Low irradiation is presented mostly between 7h-9h and 16h-18h, as it was expected. Medium irradiation among 9h-11h and 2h-4h, and finally, the highest irradiation is obtained at noon. From this, the data in blue in figure 44 which represented the hours with the lowest irradiation were used in the construction of the isoreliability curves to compare the sizing results with the ones given in daily scale.

Therefore, figure 45 shows the hourly isoreliability curves for each city with LLP of 0,01, 0,05 and 0,1; considering the load profile N° 1 in [108], and based on this plot, the optimal values for each curve are described in table 45 (under the same conditions that table 43).

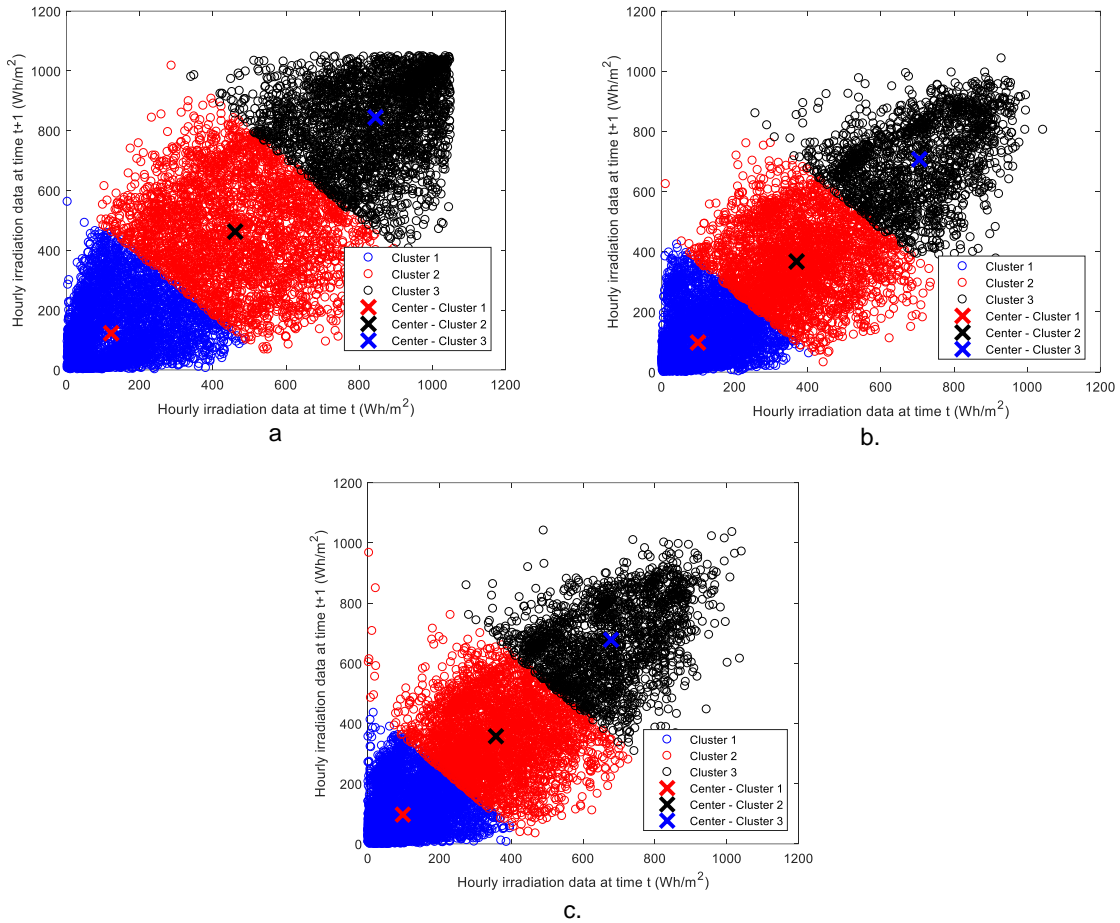
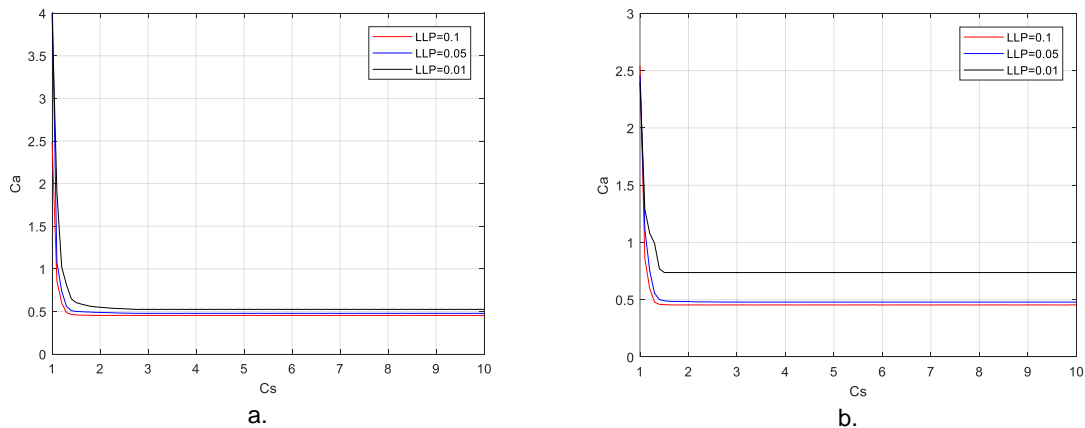
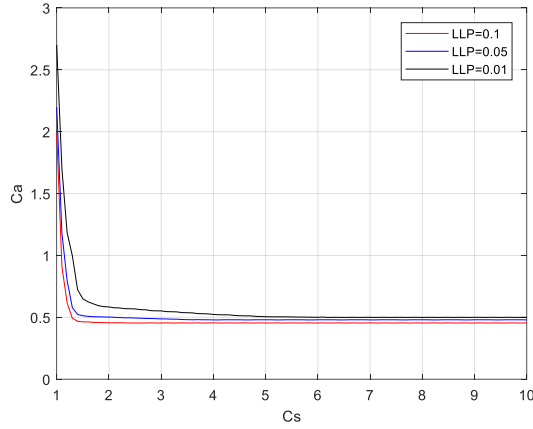


Figure 44. Space phase reconstitution clustered for hourly solar radiation: a. Cúcuta. b. Pamplona. c. Herrán.

Taking into account the more efficient sizing (LLP=0,01), a comparison between the daily and hourly results is performed and shown in figure 46. Considering this figure, we can observe the improvement of the isoreliability curve in hourly scale with respect to the daily one in each one of the cases. In a similar way to [108], the results allow us to establish that the daily analysis is not always the best choice, and that the hourly evaluation could define a better sizing in a specific place, with the advantages of requiring a lower amount of data. Improvements of 35 %, 27 % and 28 % from the daily analysis with respect to the hourly one, for a  $C_s = 1,5$  onward in Cúcuta, Pamplona and Herrán, respectively, represents a significant reduction in the sizing of the system.





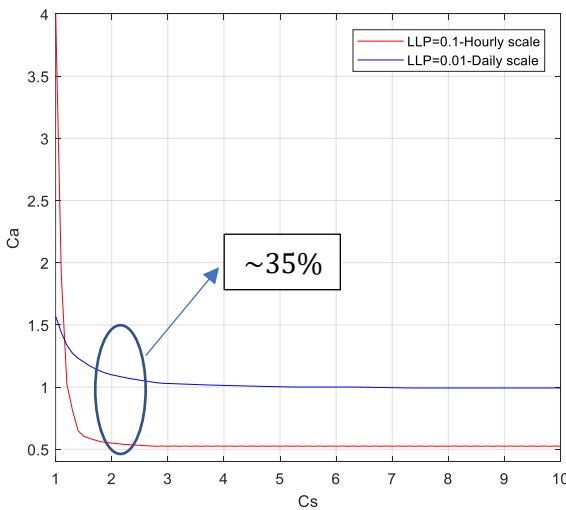
c.

Figure 45. Iso-reliability curves for hourly analysis: a. Cúcuta. b. Pamplona. c. Herrán.

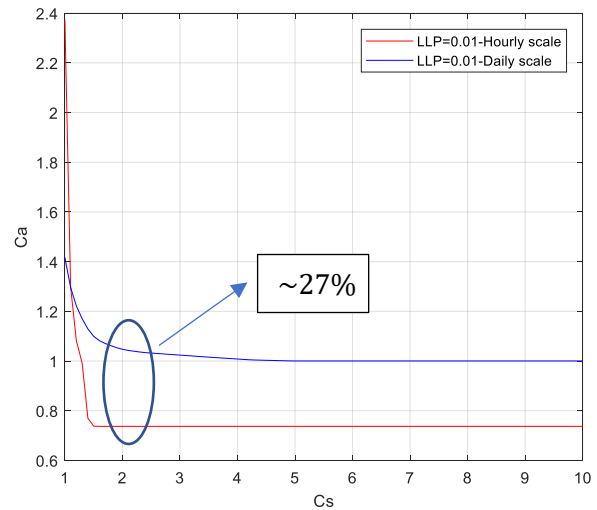
LLP	Cúcuta		Pamplona		Herrán	
	$C_A$	$C_S$	$C_A$	$C_S$	$C_A$	$C_S$
0,01	0,65	1,4	0,73	1,4	0,72	1,4
0,05	0,56	1,3	0,5	1,32	0,57	1,3
0,1	0,49	1,27	0,46	1,27	0,49	1,25

Table 45. Optimal pairs  $C_A, C_S$  for different LLP values in daily scale analysis.

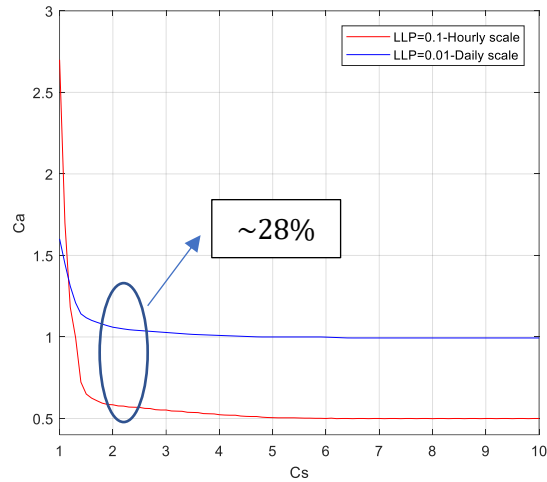
Thus, the ANN and ANFIS models acquire an advantage regarding to the other information sources indicated in the previous section, by allowing the acquisition of updated data for the application of sizing LLP methods in a hourly scale. In order to indicate the accuracy of the AI models with respect to the measured data, a comparison for the isoreliability curve with a LLP=0,01 is performed in figure 47 for each city, where are also included the data from PVGIS. It is important to note that the data from PVGIS only has the same period of evaluation in the city of Cúcuta, i.e., data until 2015. For the other cities, the curves for data measured and AI models are updated until 2017, which can represent a disadvantage for the curve from PVGIS data.



a.

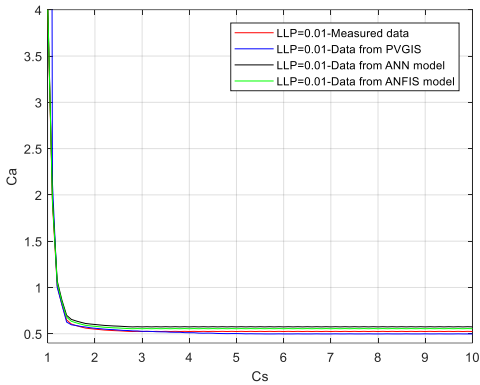


b.



c.

Figure 46. Comparison between hourly and daily LLP method with a LLP=0,01: a. Cúcuta. b. Pamplona. c. Herrán.



a.

b.

c.

c.

c.

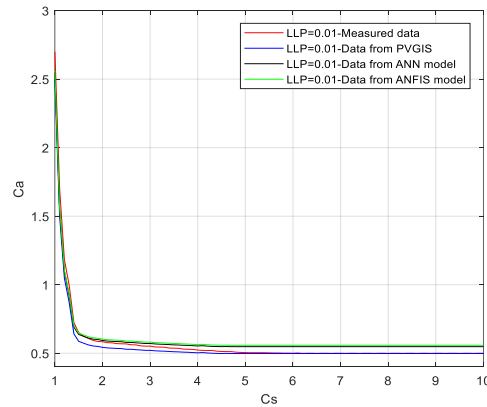


Figure 47. Comparison in hourly scale for the sizing with a LLP=0,01: a. Cúcuta. b. Pamplona. c. Herrán.

Figure 47 shows that the AI models fit the curve of measured data with high accuracy (with an  $R^2 \cong 0,99$ ) for the three cities. The results for the data from PVGIS present a low  $R^2$  for the three cities due to the large difference that it shows for a  $C_S = 1$  where the  $C_A$  value can reach up to three times the reference value (measured data); deleting this value, the adjustment improves considerably to the similar values achieved for the AI models, except for the city of Pamplona where the  $R^2$  obtained was of 0,937 (even without taking into account the first value).

# Chapter 6

## Expansion of the models

AI models were implemented in three cities of the Department of Norte de Santander considering that in these locations there were irradiation data available for the training of the networks. Thus, with specific climatological data acting as inputs and their corresponding irradiation values different links were constructed which later, with only the input variables, could estimate the irradiation profile with a high reliability. Therefore, since the places under analysis have measured solar information, the main objective of the implementation of the models was to verify their behavior and guarantee that this type of structures can be used as solar information sources with a high accuracy in other places. Despite of the above, additional benefits in the cities evaluated can be obtained from the proposed AI models such as: updated data (for instance for the city of Cúcuta where irradiation values are accessible until the year 2015, while other climatological variables are available until 2018), adaptation of this type of models in photovoltaic management system for analysis of grid integration or system information for easier access that the current ones offered by the IDEAM.

Since the benefits of the AI models mentioned above are out of the focus of this work, their implementation in the cities of Cúcuta, Pamplona and Herrán is used only as a reference for their expansion and application in other places of Norte de Santander where irradiation data are not directly measured.

To expand the models to other places in Norte de Santander that lack irradiation measurements, the following comparison process is performed: the model for the city of Herrán is evaluated with input data from the meteorological stations of the cities of Cúcuta and Pamplona; this considering that the cities of Herrán and Pamplona have similar geographical and weather conditions, hence, the model of Herrán can be analyzed in other location with similar conditions (Pamplona) and in another location with very different characteristics (Cúcuta). These scenarios are exposed in table 46 where the main geographical characteristics and the limits of the climatological variables required by the models are shown for the three cities.

City	Altitude	Latitude	Longitude	Temperature		Humidity		Wind speed		Sunshine	
				Min	Max	Min	Max	Min	Max	Min	Max
Cúcuta	311	7,898	-72,487	15,3	38,8	27	99	2e-16	7	0	1
Pamplona	2362	7,360	-72,667	7,5	21,5	19	100	0,083	7	0	1
Herrán	2040	7,506	-72,485	8	28,6	29	100	0,067	7	0	1

Table 46. Main geographical characteristics and climate conditions for the cities of Cúcuta, Pamplona and Herrán.

In a similar way to the one exposed in section 4.2.2, a 20 % of the total data was used to assess the Herrán's model from the information provided by the IDEAM for the cities of Cúcuta and Pamplona, obtaining the results displayed in table 47. In this table several error indicators of the Herrán's model are compared with the ones calculated for the models trained with data of the same city.

The analysis of table 47 shows that the Herrán's model was implemented with success in the city of Pamplona with very small error variations; on the other hand, the results in the city of Cúcuta demonstrated variations with respect to the own model of up to 21,7 % and 23 % for the RMSE and

MAPE indicator, respectively, although they do not represent a very large deviation, they could represent a significant loss of accuracy in the sizing process.

	Cúcuta		Pamplona	
	Own model	Hérran model	Own model	Hérran model
<b>RMSE (Wh/m<sup>2</sup>)</b>	173,8	211,53	117,6	120,34
<b>MAE (Wh/m<sup>2</sup>)</b>	130,2	148,97	80,1	82,46
<b>MAPE</b>	21,3	26,2	19,7	19,85
<b>R<sup>2</sup></b>	0,887	0,7594	0,9078	90,24
<b>NRMES</b>	0,163	0,2018	0,1107	0,1156

Table 47. Comparison of the error indicators for the Herrán’s model in other cities.

For the simulation of the data with the Herrán’s model, the own normalization range in each case were applied, i.e., the normalization process used in the original model for Cúcuta and Pamplona was maintained since adapting the climatological data from these cities to the normalization range for Herrán led to greater errors. This is particularly important because it can be inferred that the data from other locations must be coupled to the structure of the model, which could decrease the reliability of the estimation.



Figure 48. Expansion of the estimation models in Norte de Santander.

The process to obtain the results in table 47 was performed again using the Pamplona’s model with the input data from Herrán and Cúcuta presenting a similar behavior to the Herrán’s model with data from Pamplona and Cúcuta. Therefore, the conclusions from this section suggest that models trained in a specific place can be implemented in other locations if they have similar geographical and weather conditions. This is very important as an outcome of this research study; based on the previous argument, we can state that for different zones of the Department of Norte de Santander which do not currently have solar irradiation measurements, but they measure temperature, sunshine, wind speed, and humidity, we can use the proposed AI models to estimate the solar radiation in a reliable way for the sizing process of PV systems.

Norte de Santander					Station recommended				
Station	Municipality	Altitude (m)	Latitude	Longitude	Station	Municipality	Altitude (m)	Latitude	Longitude
Ábrego Centro Administrativo	Ábrego	1430	8,0872	-73,2231	Metromedellín	Medellín, Antioquia	1456	6,33	-75,55
Ragonvalia	Ragonvalia	1550	7,5766	-72,4838					
La Playa	La Playa	1500	8,2175	-73,235					
Aguas Claras	Ocaña	1430	8,3152	-73,3575					
Tibú	Tibú	50	8,6383	-72,7267	San Marcos	San Marcos, Sucre	27	8,6	-75,14
Salazar	Salazar	860	7,7746	-72,8306	Villeta	Villeta, Cundinamarca	880	5,02	-74,47
Silos	Silos	2765	7,2075	-72,753	San Cayetano	San Cayetano, Cundinamarca	2807	5,33	-74,02
Escuela Agronómica CÁCHIRA	Cáchira	1882	7,7352	-73,0516	Villamaría	Villamaría, Caldas	1898	5,05	-75,51
Instituto Agronómico Convención	Convención	1076	8,4705	-73,3438	Maceo	Maceo, Antioquia	1112	6,57	-74,79
Risaralda*	El Zulia	90	8,233	-72,533	Vizcaina-La Lizama	Barrancabermeja, Santander	129	6,98	-73,71

\*Station suspended currently.

Table 48. Stations recommended for the construction of irradiation estimation models with application in Norte de Santander.

Considering the previous conclusion, table 48 relates the different meteorological stations which measure the four variables used as inputs in the AI models for Norte de Santander and several meteorological stations in other places of the country that could be used for the creation of the model for the first stations due to the similarity of their geographical and weather conditions. In this table, despite of that the Risaralda station in El Zulia is suspended, it is included in the analysis considering its possible activation in the future.

With the information in table 48, a map showing weather stations from different places in Colombia with solar information matching the weather conditions of places within Norte de Santander is shown in figure 48. In this map, the stations recommended in table 48 are associated with the geographical points of Norte de Santander where could be used. Therefore, the present research proposes a new solar information source with application in different zones of the Department with a high reliable level for the sizing of PV systems as a tool for the penetration of this technology in the next years.



## Chapter 7

# Conclusions and recommendations

Irradiation indirect estimation models have been developed for three cities of the Department of Norte de Santander (Colombia), using AI techniques (ANN and ANFIS) from climatological variables of easy access in the region (humidity, temperature, wind speed and sunshine) in hourly, daily and monthly scale. A first performance analysis showed that the models have the expected behavior obtaining estimation errors lower than several works in the literature mainly in the highest time scale, demonstrating that improvements in the results of each model can be achieved by increasing the estimation period. Between the two AI proposed models, it was found that ANFIS offers better performance with respect to ANN, although the difference is not very significant to categorize the ANFIS model as the best alternative for the estimation process.

In addition to this, ANN and ANFIS models were evaluated together with other solar information sources (satellite and empirical ones) in specific sizing cases of PV systems in order to identify the benefits of these models in this field. The AI models performed estimations very close to the reference in all the scenarios, unlike the rest of the databases analyzed, which according to the time scale required in the sizing method, presented large variations in the accuracy of the results. This represents an important advantage for the deployment of this type of models in the region, because not only these models can be used in several sizing methods that require different time scales (which is not common for all information sources since only one of them has data available in all scales) but also because their application guarantees a high level of precision and reliability.

It is worth to indicate that the goal of this research study is not to verify if the available solar information sources are accurate or not (since a more extensive and deep study about it is needed and also considering that this document assumed that the data from IDEAM are the closest to the real values of interest); in a similar way, the final purpose is not to recognize the best sizing method for a PV system. The main objective is to analyze the impact of an alternative solar information source based on AI techniques; thus, considering this approach and the obtained results, we can conclude that the AI proposed models are a remarkable tool for the description of solar radiation profiles in Norte de Santander if these models are trained with enough reliable data, and therefore, these can exert a positive influence in the implementation of PV systems.

Besides, after the validation of the three developed models, we found that their application is not limited just to the places where the models were trained, but also to zones with similar weather and geographical conditions; thus, models constructed from different regions of Colombia could be adapted for predicting irradiation data with a high accuracy in cities of Norte de Santander where this variable is not directly measured (as table 48 suggests). The above can be used as a strategy to mitigate the negative effects of lacking accurate information of this variable on the growth of the PV technology in the country; effects that were identified by the government of Colombia in the last years.

Finally, we recommend for future works to broaden the scope of the model, i.e., to test the model in many more locations to define the maximum and minimum geographical and weather conditions that the prediction can tolerate; in this sense, it could be possible to cover a greater number of areas with the same model and reduce the list displayed in table 48. Further, there are models in the literature that combine training data from more than one station to extend the application range; this option could be also performed and compared with the initial recommendation to find the choice with the best results for the estimation. This could be done not only in terms of performance but also based on complexity, computational cost and reduction of resources for its realization.

APPENDIX A

<i>t</i>	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,8	1,6	2,39	3,19	3,99	4,78	5,58	6,38	7,17
0,1	7,97	8,76	9,55	10,34	11,13	11,92	12,71	13,5	14,28	15,07
0,2	15,85	16,63	17,41	18,19	18,97	19,74	20,51	21,28	22,05	22,82
0,3	23,58	24,34	25,1	25,86	26,61	27,37	28,12	28,86	29,61	30,35
0,4	31,08	31,82	32,55	33,28	34,01	34,73	35,45	36,17	36,88	37,59
0,5	38,29	38,99	39,69	40,39	41,08	41,77	42,45	43,13	43,81	44,48
0,6	45,15	45,81	46,47	47,13	47,78	48,43	49,07	49,71	50,35	50,98
0,7	51,61	52,23	52,85	53,46	54,07	54,67	55,27	55,87	56,46	57,05
0,8	57,63	58,21	58,78	59,35	59,91	60,47	61,02	61,57	62,11	62,65
0,9	63,19	63,72	64,24	64,76	65,28	65,79	66,29	66,8	67,29	67,78
1	68,27	68,75	69,23	69,7	70,17	70,63	71,09	71,54	71,99	72,43
1,1	72,87	73,3	73,73	74,15	74,57	74,99	75,4	75,8	76,2	76,6
1,2	76,99	77,37	77,75	78,13	78,5	78,87	79,23	79,59	79,95	80,29
1,3	80,64	80,98	81,32	81,65	81,98	82,3	82,62	82,93	83,24	83,55
1,4	83,85	84,15	84,44	84,73	85,01	85,29	85,57	85,84	86,11	86,38
1,5	86,64	86,9	87,15	87,4	87,64	87,89	88,12	88,36	88,59	88,82
1,6	89,04	89,26	89,48	89,69	89,9	90,11	90,31	90,51	90,7	90,9
1,7	91,09	91,27	91,46	91,64	91,81	91,99	92,16	92,33	92,49	92,65
1,8	92,81	92,97	93,12	93,28	93,42	93,57	93,71	93,85	93,99	94,12
1,9	94,26	94,39	94,51	94,64	94,76	94,88	95	95,12	95,23	95,34
2	95,45	95,56	95,66	95,76	95,86	95,96	96,06	96,15	96,25	96,34
2,1	96,43	96,51	96,6	96,68	96,76	96,84	96,92	97	97,07	97,15
2,2	97,22	97,29	97,36	97,43	97,49	97,56	97,62	97,68	97,74	97,8
2,3	97,86	97,91	97,97	98,02	98,07	98,12	98,17	98,22	98,27	98,32
2,4	98,36	98,4	98,45	98,49	98,53	98,57	98,61	98,65	98,69	98,72
2,5	98,76	98,79	98,83	98,86	98,89	98,92	98,95	98,98	99,01	99,04
2,6	99,07	99,09	99,12	99,15	99,17	99,2	99,22	99,24	99,26	99,29
2,7	99,31	99,33	99,35	99,37	99,39	99,4	99,42	99,44	99,46	99,47
2,8	99,49	99,5	99,52	99,53	99,55	99,56	99,58	99,59	99,6	99,61
2,9	99,63	99,64	99,65	99,67	99,68	99,69	99,7	99,71	99,72	
3	99,73									
3,5	99,95									
4	99,994									
4,5	99,9993									

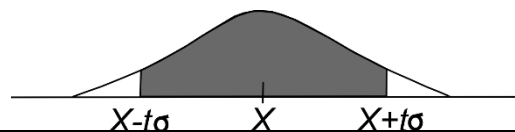


Table A1. The percentage probability,  $Prob(\text{without } t\sigma) = \int_{X-t\sigma}^{X+t\sigma} G_{X,\sigma}(x)dx$  as a function of  $t$ . Appendix B in [81].

## APPENDIX B

The most notable goodness of fit tests are the *chi square*, *Cramér–Von Mises*, *Kolmogorov–Smirnov*, *Shapiro–Wilk* and *Anderson–Darling* tests; but the Anderson–Darling test is currently considered to be superior [112], being one of the most powerful tools when normality is tested [113]. Therefore, this method is the selected one for the verification process.

As a test of goodness of fit, Anderson Darling test is used to determine whether sample belongs to a certain distribution by calculating the Anderson-Darling statistic between cumulative distribution function (CDF) and empirical probability density function (EDF) [114]. Let  $x_1, \dots, x_N$  be the  $N$  observations of the given sample, whose *normality test* based on the *Anderson-Darling* method follows the procedure below.

First, the set of ranges are sorted in ascending order:

$$x_1 \leq x_2 \leq x_3 \dots, x_N$$

and, since the possible underlying normal distribution is unknown, the mean and variance are estimated, being the first one the average of the data and the second one calculated from equation 2. With  $F(x)$  being the standard normal CDF as:

$$F(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x - \hat{x}}{\sqrt{2}\sigma}\right)$$

Considering that the error function is defined by:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

and using the standardized sample or z-scores from equation 3, the Anderson-Darling statistic  $A^2$  can be computed as:

$$A^2 = - \sum_{i=1}^N \frac{2i-1}{N} [\ln F(z_i) + \ln(1 - F(z_{N+1-i}))] - N$$

where  $z_i$  represents the corresponding z-scores of each measurements or sample in the set. Then, the *null hypothesis*  $H_0$  (in inferential statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups; for this case, the hypothesis are  $H_0$  which assumes that the sample is normally distributed and  $H_1$  which defined the contrary), can be rejected if the modified statistic exceeds a given threshold:

$$A^{2*} \triangleq A^2 \left(1 + \frac{4}{N} - \frac{25}{L^2}\right) > \gamma_\alpha$$

The threshold  $\gamma_\alpha$  is fixed for a chosen level of significance  $\alpha$ , where  $0 \leq \alpha \leq 1$  is defined as:

$$\alpha = P\{A^{2*} > \gamma_\alpha \mid H_0\}$$

that is, the probability of rejecting the null hypothesis while true. It can be obtained numerically by Monte Carlo simulations, although equation 8 shows an expression for its calculation [115]:

$$\gamma_\alpha = a_\alpha \left(1 + \frac{b_0}{N} + \frac{b_1}{N^2}\right)$$

in which values of  $a_\alpha$ ,  $b_0$  and  $b_1$  are given in table B1 for a prescribed significance level  $\alpha$ .

Significant level $\alpha$	$a_\alpha$	$b_0$	$b_1$
0,2	0,5091	-0,756	-0,39
0,1	0,6305	-0,75	-0,8
0,05	0,7514	-0,795	-0,89
0,025	0,8728	-0,881	-0,94
0,01	1,0348	-1,013	-0,93
0,005	1,1578	-1,063	-1,34

Table B1. Coefficients to obtain the critical values for Anderson-Darling Goodness of Fit Test.

The detection probability can then be straightforwardly computed as:

$$P_d = P\{A^{2*} > \gamma_\alpha \mid H_0\} = 1 - \alpha$$

Therefore, by means of the *Anderson-Darling method* each one of the sets (groups of 12 values which represent the evaluation range by day) of the variables were assessed with a level of significance  $\alpha = 0,05$  as in [79], with which a critical value  $\gamma_\alpha = 0,697$  was obtained, following the process presented in a summarized form below.

*Algorithm:* Anderson-Darling test for Normality of ranges between node  $i$  and  $j$

**Require:**  $\{\hat{d}^{(l)}_{i,j}\}_{l=1}^L, \alpha$

- 1: Sort ranges in ascending order:  $\hat{d}^{(1)}_{i,j} \leq \hat{d}^{(2)}_{i,j} \leq \dots \leq \hat{d}^{(L)}_{i,j}$ .
- 2: Estimate mean and standard deviation of the sample.
- 3: Compute the standardized sample  $\left\{ \omega^{(l)} = \frac{\hat{d}^{(1)}_{i,j} - \hat{\mu}}{\sigma} \right\}_{l=1}^L$
- 4: Evaluate  $A^2$  and  $A^{2*}$  statistics.
- 5: **if**  $A^{2*} \geq \gamma_\alpha$  **then**
- 6: Reject the null hypothesis: the sample is not normally distributed.
- 7: **else**
- 8: Accept the null hypothesis: the sample is normally distributed.
- 9: **end if**

As a complement, a *t-test* is applied to confirm the assumption in terms of the type of distribution of the sample [78]. For this case, a *two-sided t-test* is implemented, which is typically applicable when there are two tails in the structure, such as in the normal distribution, and correspond to considering either direction significant. The terminology "tail" is used because the extreme portions of distributions, where observations lead to rejection of the *null hypothesis*, are small and often "tail off" toward zero as in the normal distribution or "bell curve". This test is supported by the equation B1 and it is illustrated in figure B1 [83].

$$-t_{\frac{\alpha}{2}, N-1} \leq \frac{\sqrt{N}(x - \hat{x})}{\sigma} \leq t_{\frac{\alpha}{2}, N-1} \quad Eq. B1$$

where  $\alpha$  is called *the level of significance of the test*, is usually set in advance, with commonly chosen values being  $\alpha = 0,1, 0,05, 0,005$  and  $t_{\frac{\alpha}{2}, N-1}$  is the 100  $\alpha/2$  upper percentile value of *t-distribution*

with  $N - 1$  degrees of freedom. In statistics, a *t-distribution* is a family of curves depending on a single parameter  $\nu$  (the degrees of freedom), where  $\nu$  can be  $\nu = 1, 2, 3, \dots$  and as it becomes larger, the *t-distribution* becomes more and more like a standard normal density [116].

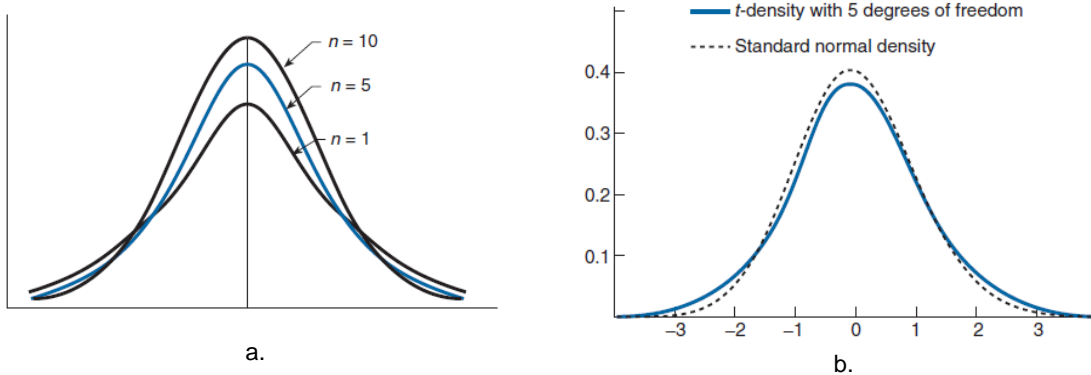


Figure B1. A *t-distribution*: a. Effect of the increase of the degrees of freedom. b. Comparison between a *t-distribution* with a *standard normal distribution*.

The probability density function (pdf) of the *t-distribution* or *Student's t distribution* is given by equation B2.

$$y = f(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}} \quad \text{Eq. B2}$$

where  $\nu$  is the degrees of freedom and  $\Gamma(\nu)$  is the Gamma function. The result  $y$  is the probability of observing a particular value of  $x$  from a *Student's t distribution* with  $\nu$  degrees of freedom.

Table A3 (from the Appendix of Tables in [83]) presents several values of  $t_{\alpha,N}$  for a wide variety of values of  $N$  and  $\alpha$ . For the assessment in this alternative method, again an  $\alpha = 0,05$  was selected and a  $t_{\frac{\alpha}{2},N-1} = 2,201$  (for a  $N = 12$ ) were used in the calculation of the verification process.

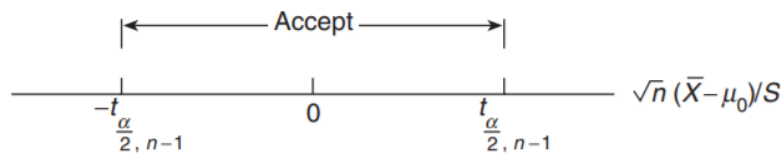


Figure B2. The two-sided t-test.

## APPENDIX C

As it was mentioned in section 4.2.1 (step 3),  $K_t$  is defined as the ratio  $H/H_{ext}$ , where  $H$  and  $H_{ext}$  are the monthly average of the daily solar irradiation and the daily extraterrestrial solar irradiation on a horizontal surface, respectively. To be able speaking about daily extraterrestrial solar radiation is needed before speaking about the *solar constant* ( $I_0$ ); it is the amount of energy coming from the Sun that perpendicularly affects a surface of unitary area placed outside the Earth's atmosphere, at an average distance of  $150 \times 10^6$  km from the Sun [117].

The value of the solar constant has been investigated; nowadays improvements are made in the instruments and in the methodologies for its determination as the average value of numerous measurements. The value currently used is  $1.370 \text{ W/m}^2$  and oscillates approximately  $1,2 \text{ W/m}^2$  between the maximum and the minimum of the cycle. This value has been adopted as a solar constant in different parts of the world, including some educational exercises at NASA. However, the value adopted as solar constant by the World Meteorological Organization (WMO) until the last calibration performed during the year 2000 is:  $\bar{I}_0 = 1.367 \text{ W/m}^2$  with an error of  $\pm 7 \text{ W/m}^2$ .

As the intensity of the solar energy varies inversely proportional to the square of the distance to the Sun, then in the translational movement of the Earth in Earth orbit changes the Earth-Sun distance during the year, causing a variation of incident extraterrestrial solar radiation on a surface normal to the solar ray, as illustrated in figure C1.

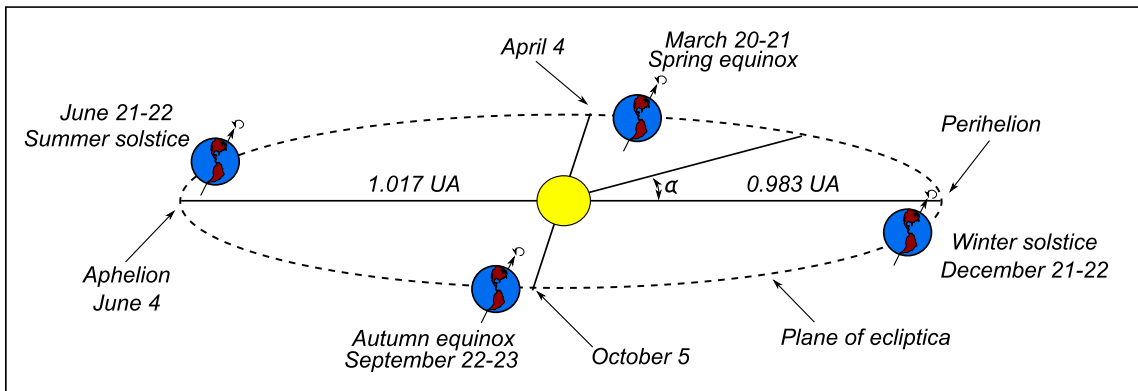


Figure C1. Movement of the Earth around Sun.

Analytically, incident extraterrestrial solar radiation can be determined by equation C1, for a day.

$$I_{nd} = \bar{I}_0 \left( \frac{R_0}{R} \right)^2 \quad \text{Eq. C1}$$

where the ratio  $R_0/R$  is a Fourier series which represents the Earth-Sun distance with a maximum error of 0,01 percent. This relationship is explained with more detailed as follows. The Earth in its movement around the Sun describes an elliptical orbit in which the average Earth-Sun distance is approximately  $149,46 \times 10^6$  km (value called an Astronomical Unit U. A.). The orbit of the Earth can be described in polar coordinates by the following expression:

$$R = \frac{a(1 - e^2)}{(1 + e \cos \alpha)} \quad \text{Eq. C2}$$

where:  $R = \text{Earth} - \text{Sun distance}$ ,  $\alpha = \text{Earth angular position in the orbit}$ ,  $e = \text{eccentricity of the Earth's orbit}$  ( $e = 0,01673$ ),  $a = \text{astronomical unit (semimajor axis of the ellipse)}$

The Earth angular position for this case is given by:

$$\alpha = \frac{2\pi(nd - 1)}{365}$$

where:  $nd = \text{number of day of the year}$

When  $\alpha = 0^\circ$  the Earth is closer to the Sun (perihelion), as it is observed in figure C1, thus from equation C2 is obtained:

$$R = a(1 - e) = 0,983 \text{ U.A.}$$

When  $\alpha = 180^\circ$ , the Earth is in the most distant position of the Sun (aphelion); in this point:

$$R = a(1 + e) = 1,017 \text{ U.A.}$$

The distance  $R$  for radiometric effects can be expressed by means of a simple computation equation, defined by Spencer in [118] as a Fourier series, with a maximum error of 0,01 percent:

$$\left(\frac{R_0}{R}\right)^2 = 1,00011 + 0,034221 \cos \alpha + 0,00128 \sin \alpha + 0,000719 \cos 2\alpha + 0,000077 \sin 2\alpha$$

where:  $R_0 = \text{Earth - Sun average distance (1 U.A.)}$

Therefore, considering above the daily solar radiation that hits a horizontal surface outside the Earth's atmosphere denoted by  $H_{ext}(nd)$ , where  $nd$  is the number of day of the year, represents the amount of energy incident on that surface from sunrise to sunset; if an atmosphere totally transparent to that radiation existed, it would arrive unaltered on the terrestrial surface and would have the same value and behavior. The following expression allows to determine it:

$$H_{ext}(n) = \int I_n \cos \theta dt$$

where:  $I_{nd} = \text{Incident extraterrestrial solar radiation for the day } n \text{ of the year}$

$\theta = \text{Angle of incidence}$

Whose solution according to the mathematical handling presented in [117], is given by:

$$H_{ext}(nd) = \frac{24}{\pi} \times \bar{I}_0 \left(\frac{R_0}{R}\right)^2 \left( \cos \phi \cos \delta \sin \omega + \frac{2\pi\omega}{360^\circ} \sin \phi \sin \delta \right) \text{ Eq. C3}$$

being:  $\phi = \text{Latitude}$ ,  $\delta = \text{Decline angle}$ ,  $\omega = \text{Hour angle}$

The variables  $\delta$  and  $\omega$  corresponding to Sun position angles according to the equatorial coordinate system. The hour angle ( $\omega$ ) is the angle formed at the pole by the intersection between the meridian of the observer and the meridian of the Sun; it is expressed in units of arc (degrees) or in units of time (hours); its conversion is: 1 hour = 15 ° and it is defined by equation C4.

$$\omega = \frac{360}{24}(t - 12) \quad \text{or} \quad \omega = \frac{2\pi}{24}(t - 12) \text{ Eq. C4}$$

where  $t$  is the local hour. Equation C4 is used for the calculation of  $\omega$  to a specif hour, but in the case of the equation C3,  $\omega$  must be defined from sunrise to sunset, whereby is determined equation C5.

$$\omega = \cos^{-1}(-\tan \phi \tan \delta) \text{ Eq. C5}$$

From which the astronomical duration of the day, i.e., the duration in hours from sunrise to sunset, can be calculated in this way:

$$N_{s-s} = \frac{2}{15} \omega \quad \text{Eq. C6}$$

Equation C6 is valid if the absolute value of  $(-\tan \phi \tan \delta) \leq 1$ . For high latitudes (greater than  $66,7^\circ$ ), where depending on the time of year that condition is not satisfied, implies that the days can have a duration equal to 24 hours, the Sun is not hidden, or equal to 0 hours, the Sun stays below the horizon, depending on the day of the year.

Now, the angle formed between the equatorial plane of the Earth and the Earth-Sun line is called the solar declination ( $\delta$ ), and it is shown in figure C2. Due to the movement of the Earth around the Sun the value of this angle varies during the year. The sign of the declination is positive (+) when the Sun strikes perpendicularly somewhere in the northern hemisphere, between March 21 (spring equinox) and September 23 (autumn equinox), and negative (-) when it incites perpendicularly over some place in the southern hemisphere, between September 23 (autumn equinox) and March 21 (spring equinox), and varies between  $-23,45^\circ$ , when the Sun is in the lowest part from the southern hemisphere (winter solstice December 21/22), and  $+23,45^\circ$ , when it is in the highest part of the northern hemisphere (summer solstice June 21/22). Two times during the year it takes zero value, when the Sun passes over the terrestrial Equator, during the equinoxes (Fig. C3). The daily values of the solar declination can be calculated with a maximum error of 0,0006 rad., by means of another formula obtained by Spencer in [118]:

$$\delta = (0,006918 - 0,399912 \cos \alpha + 0,070257 \sin \alpha - 0,006758 \cos 2\alpha + 0,000907 \sin 2\alpha - 0,002697 \cos 3\alpha + 0,00148 \sin 3\alpha) \left(\frac{180}{\pi}\right)$$

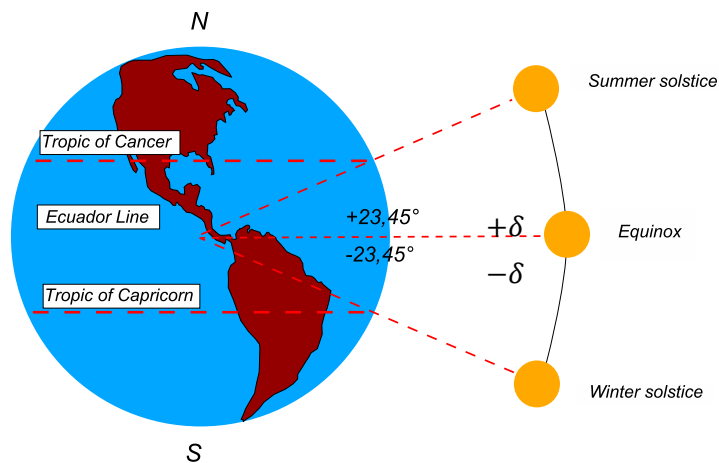


Figure C2. Diagram of the change of declination with movement of the Sun with respect to the Ecuador plane.

Other expressions of equation C3 can be found in [119] and [120] which follow the same principle but handling other approximations for the calculation of the decline angle or other interpretation in the form to present the values of the different angles in the equation.

In this way, based on equation C3, the extraterrestrial solar radiation was calculated for each day of the year, and this value was compared with each experimental value provided by the IDEAM, determining as an atypical measurement, those whose value was higher than the result from equation C3.



# References

- [1] IDEAM, "ATLAS DE RADIACIÓN SOLAR, ULTRAVIOLETA Y OZONO DE COLOMBIA - Aspectos teóricos," 2018. [Online]. Available: <http://atlas.ideam.gov.co/basefiles/5.Aspectos-teoricos.pdf>. [Accessed June 2018].
- [2] Unidad de Planeación Minero Energética (UPME), " Estudio: Integración de las energías renovables no convencionales en Colombia," 2015. [Online]. Available: <http://www1.upme.gov.co/Paginas/Estudio-Integraci%C3%B3n-de-las-energ%C3%ADas-renovables-no-convencionales-en-Colombia.aspx>. [Accessed June 2018].
- [3] Unidad de Planeación Minero Energética - UPME, "Boletín Estadístico de Minas y Energía 2018," November 2018. [Online]. Available: [http://www1.upme.gov.co/PromocionSector/SeccionesInteres/Documents/Boletines/Boletin\\_Estadistico\\_2018.pdf](http://www1.upme.gov.co/PromocionSector/SeccionesInteres/Documents/Boletines/Boletin_Estadistico_2018.pdf) [Accessed December 2018].
- [4] K. Chiteka and C. C. Enweremadu, "Prediction of global horizontal solar irradiance in Zimbabwe using artificial neural networks," *Journal of Cleaner Production*, vol. 135, pp. 701-711, 2016.
- [5] IDEAM, "Catálogo Nacional de Estaciones del IDEAM," 2018. [Online]. Available: <https://www.datos.gov.co/en/Ambiente-y-Desarrollo-Sostenible/Cat-logo-Nacional-de-Estaciones-del-IDEAM/hp9r-jxuu>
- [6] IDEAM, "Atlas de Radiación Solar, Ultravioleta y Ozono de Colombia," 2018. [Online]. Available: <http://atlas.ideam.gov.co/basefiles/Anexo-Lista-de-estaciones-convencionales-de-radiacion-global-del-Ideam.pdf>. [Accessed June 2018].
- [7] A. Jakhrani, A. K. Othman, A. R. H. Rigit and S. R. Samo, "A simple method for the estimation of global solar radiation from sunshine hours and other meteorological parameters," in *IEEE International Conference on Sustainable Energy Technologies (ICSET)*, Kandy, Sri Lanka, 2010.
- [8] A. Teke, H. Yıldırım and ÖzgürÇelik, "Evaluation and performance comparison of different models for the estimation of solar radiation," *Renewable and Sustainable Energy Reviews*, vol. 50, p. 1097–1107, 2015.
- [9] L. Wang, O. Kisi, M. Zounemat-Kermani and G. A. Salazar, "Solar radiation prediction using different techniques: model evaluation and comparison," *Renewable and Sustainable Energy Reviews*, vol. 61, p. 384–397, 2016.
- [10] S. Mekhilef, R. Saidur and M. Kamalisarvestani, "Effect of dust, humidity and air velocity on efficiency of photovoltaic cells," *Renewable and Sustainable Energy Reviews*, vol. 16, p. 2920– 2925, 2012.
- [11] B. A. L. Gwandu and D. J. Creasey, "Humidity: A factor in the appropriate positioning of a photovoltaic power station," *Renewable Energy*, vol. 6, no. 3, pp. 313-316, 1995.
- [12] S. Ghazi and K. Ip, "The effect of weather conditions on the efficiency of PV panels in the southeast of UK," *Renewable Energy*, vol. 69, p. 50e59, 2014.
- [13] F. Reis, C. Guerreiro, F. Batista, T. Pimentel, M. Pravettoni, J. Wemans, G. Sorasio and M. C. Brito, "Modelling the Effects of Inhomogeneous Irradiation and Temperature Profile on CPV Solar Cell Behavior," *IEEE JOURNAL OF PHOTOVOLTAICS*, vol. 5, no. 1, pp. 112-122, 2015.
- [14] J. Zhang, L. Zhao, S. Deng, W. Xu and Y. Zhang, "A critical review of the models used to estimate solar radiation," *Renewable and Sustainable Energy Reviews*, vol. 70, p. 314–329, 2017.
- [15] J.-G. Kim, D.-H. Kim, W.-S. Yoo, J.-Y. Lee and Y. B. Kim, "Daily prediction of solar power generation based on weather forecast information in Korea," *IET Renewable Power Generation*, vol. 11, no. 10, pp. 1268-1273, 2017.
- [16] U. K. Das, K. S. Teya, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. V. Deventer, B. Horan and A. Stojcevski, "Forecasting of photovoltaic power generation and model optimization: A review," *Renewable and Sustainable Energy Reviews*, vol. 81, p. 912–928, 2018.

- [17] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. Martinez-de-Pison and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, p. 78–111, 2016.
- [18] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin and Z. Hu, "Photovoltaic and Solar Power Forecasting for Smart Grid Energy Management," *CSEE JOURNAL OF POWER AND ENERGY SYSTEMS*, vol. 1, no. 4, pp. 38-46, 2015.
- [19] K. Chen, Z. He, K. Chen, J. Hu and J. He, "Solar energy forecasting with numerical weather predictions on a grid and convolutional networks," in *IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Beijing, China, 2017.
- [20] J. Henga, J. Wanga, L. Xiao and H. Lu, "Research and application of a combined model based on frequent pattern growth algorithm and multi-objective optimization for solar radiation forecasting," *Applied Energy*, vol. 208, p. 845–866, 2017.
- [21] R. Huang, T. Huang, R. Gadh and N. Li, "Solar Generation Prediction using the ARMA Model in a Laboratory-level Micro-grid," in *IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, Tainan, Taiwan, 2012.
- [22] R. Perdomo, E. Banguero and G. Gordillo, "STATISTICAL MODELING FOR GLOBAL SOLAR RADIATION FORECASTING IN BOGOTÁ," in *IEEE Photovoltaic Specialists Conference*, Honolulu, HI, USA, 2010.
- [23] E. G. Kardakos, M. C. Alexiadis, S. I. Vagropoulos, C. K. Simoglou, P. N. Biskas and A. G. Bakirtzis, "Application of time series and artificial neural network models in short-term forecasting of PV power generation," in *48th International Universities' Power Engineering Conference (UPEC)*, Dublin, Ireland, 2013.
- [24] C. Yang, A. A. Thatte and L. Xie, "Multitime-Scale Data-Driven Spatio-Temporal Forecast of Photovoltaic Generation," *IEEE TRANSACTIONS ON SUSTAINABLE ENERGY*, vol. 6, no. 1, pp. 104-112, 2015.
- [25] S. D. Campbell and F. X. Diebold, "Weather Forecasting for Weather Derivatives," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 6-16, 2005.
- [26] Y. Li, Y. Su and L. Shu, "An ARMAX model for forecasting the power output of a grid connected photovoltaic system," *Renewable Energy*, vol. 66, pp. 78-89, 2014.
- [27] M. B. b. M. G. b. A. R. Emanuel Federico Alsina a, "Artificial neural network optimisation for monthly average daily global solar radiation prediction," *Energy Conversion and Management*, vol. 120, p. 320–329, 2016.
- [28] H. B. Yıldırima, Ö. Çelik, A. Teke and B. Barutçu, "Estimating daily Global solar radiation with graphical user interface in Eastern Mediterranean region of Turkey," *Renewable and Sustainable Energy Reviews*, vol. 82, p. 1528–1537, 2018.
- [29] P. Neelamegama and V. A. Amirtham, "Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms," *Journal of Applied Research and Technology*, vol. 14, p. 206–214, 2016.
- [30] A. Qazi, H. Fayaz, A. Wadi, R. G. Raj, N. Rahim and W. A. Khan, "The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review," *Journal of Cleaner Production*, vol. 104, pp. 1-12, 2015.
- [31] M. Bou-Rabeea, S. A. Sulaimanb, M. S. Salehc and S. Marafid, "Using artificial neural networks to estimate solar radiation in Kuwait," *Renewable and Sustainable Energy Reviews*, vol. 72, p. 434–438, 2017.
- [32] M. H. Alobaidi, P. R. Marpu, T. B. M. J. Ouarda and H. Ghedira, "Mapping of the Solar Irradiance in the UAE Using Advanced Artificial Neural Network Ensemble," *IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING*, vol. 7, no. 8, pp. 3668-3680, 2014.
- [33] M. Rizwan, M. Jamil, S. Kirmani and D. Kothari, "Fuzzy logic based modeling and estimation of global solar energy using meteorological parameters," *Energy*, vol. 70, pp. 685-691, 2014.
- [34] S. Chen, H. Gooi and M. Wang, "Solar radiation forecast based on fuzzy logic and neural networks," *Renewable Energy*, vol. 60, p. 195e201, 2013.
- [35] R. C. Deo, X. Wenb and F. Qi, "A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset," *Applied Energy*, vol. 168, p. 568–593, 2016.

- [36] S. Belaid and A. Mellit, "Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate," *Energy Conversion and Management*, vol. 118, p. 105–118, 2016.
- [37] S. Shamshirband, K. Mohammadi, H. Khorasanizadeh, P. L. Yee, M. Lee, D. Petković and E. Zalnezhad, "Estimating the diffuse solar radiation using acoupled support vector machine–wavelet transform model," *Renewable and Sustainable Energy Reviews*, vol. 56, p. 428–435, 2016.
- [38] A. Aybar-Ruiz, S. Jiménez-Fernández, L. Cornejo-Bueno and C. Casanova-Mateo, "A novel Grouping Genetic Algorithm–Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs," *Solar Energy*, vol. 132, p. 129–142, 2016.
- [39] M. Paulescu, E. Paulescu, P. Gravila and V. Badescu, "Satellite Based Models for Deriving - Online Available Database," in *Weather Modeling and Forecasting of PV Systems Operation*, London, UK, Springer, 2013, pp. 29-35.
- [40] J. A. Ruiz-Ariasa and C. A. Gueymard, "Worldwide inter-comparison of clear-sky solar radiation models: Consensus-based review of direct and global irradiance components simulated at the earth surface," *Solar Energy*, vol. 168, p. 10–29, 2018.
- [41] A. Ayeta and P. Tandeo, "Nowcasting solar irradiance using an analog method and geostationary satellite images," *Solar Energy*, vol. 164, p. 301–315, 2018.
- [42] E. Gerdali, F. Romano and E. Ricciardelli, "An Advanced Model for the Estimation of the Surface Solar Irradiance Under All Atmospheric Conditions Using MSG/SEVIRI Data," *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, vol. 50, no. 8, pp. 2934-2953, 2012.
- [43] C. Bertrand, G. Vanderveken and M. Journée, "Evaluation of decomposition models of various complexity to estimate the direct solar irradiance over Belgium," *Renewable Energy*, vol. 74, pp. 618-626, 2015.
- [44] T. Hove, E. Manyumbu and G. Rukweza, "Developing an improved global solar radiation map for Zimbabwe through correlating long-term ground- and satellite-based monthly clearness index values," *Renewable Energy*, vol. 63, pp. 687-697, 2014.
- [45] D. Palmer, E. Koubli, I. Cole, T. Betts and R. Gottschalg, "Satellite or ground-based measurements for production of site specific hourly irradiance data: Which is most accurate and where?," *Solar Energy*, vol. 165, p. 240–255, 2018.
- [46] F. Besharat, A. A. Dehghan and A. R. Faghii, "Empirical models for estimating global solar radiation: A review and case study," *Renewable and Sustainable Energy Reviews*, vol. 21, p. 798–821, 2013.
- [47] M. Despotovic, V. Nedic, D. Despotovic and S. Cvetanovic, "Review and statistical analysis of different global solar radiation sunshine models," *Renewable and Sustainable Energy Reviews*, vol. 52, p. 1869–1880, 2015.
- [48] A. I. Ibrahim and T. Khatib, "A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm," *Energy Conversion and Management*, vol. 138, p. 413–425, 2017.
- [49] R. Azimi, M. Ghayekhloo and M. Ghofrani, "A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting," *Energy Conversion and Management*, vol. 118, p. 331–344, 2016.
- [50] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petkovic and C. Sudheer, "A support vector machine–firefly algorithm-based model for global solar radiation prediction," *Solar Energy*, vol. 115, p. 632–644, 2015.
- [51] S. Sobri, S. Koohi-Kamali and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Conversion and Management*, vol. 156, p. 459–497, 2018.
- [52] P. Ponce Cruz, "Capítulo 3: Redes Neuronales Artificiales," in *Inteligencia Artificial con Aplicaciones a la Ingeniería*, Ciudad de México D.F., México, Alfaomega, 2010, pp. 193-282.
- [53] V. H. Queja, J. Almoroxa, J. A. Arnaiz and L. Saito, "ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment," *Journal of Atmospheric and Solar–Terrestrial Physics*, vol. 155, no. 1, p. 62–70, 2017.

- [54] Y. Kashyap, A. Bansal and A. K. Sao, "Solar radiation forecasting with multiple parameters neural networks," *Renewable and Sustainable Energy Reviews*, vol. 49, p. 825–835, 2015.
- [55] J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665 - 685, 1993.
- [56] Google, "Google Maps (Route Cúcuta - Pamplona)," 2018. [Online]. Available: <https://www.google.es/maps/dir/Pamplona,+Norte+de+Santander,+Colombia/C%C3%BAcuta,+Norte+de+Santander,+Colombia/@7.6319206,-72.8457572,10z/am=t/data=!4m14!4m13!1m5!1m1!1s0x8e68811722aa0c15:0x83b790e98f7db7d4!2m2!1d-72.6503369!2d7.37823!1m5!1m1!1s0x8e66459>. [Accessed August 2018].
- [57] The World Bank, "Globar Solar Atlas (GSA)," Solargis, 2018. [Online]. Available: <http://globalsolaratlas.info/about>. [Accessed June 2018].
- [58] NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program, "POWER Project Data Sets," 2018. [Online]. Available: <https://power.larc.nasa.gov/>.
- [59] European Commission Joint Research Centre (JRC), "Overview of PVGIS data sources and calculation methods," 2018. [Online]. Available: [http://re.jrc.ec.europa.eu/pvg\\_static/methods.html](http://re.jrc.ec.europa.eu/pvg_static/methods.html).
- [60] Grupo de Investigación y Desarrollo en Microelectrónica Aplicada (GIDMA), "ANÁLISIS DE LA OFERTA DE ENERGÍA SOLAR FOTOVOLTAICA EN ZONAS RURALES DE NORTE DE SANTANDER," Unidad de Planeación Minero Energética (UPME), Cúcuta, Colombia, 2017. Available: [http://persnds.ufps.edu.co/pers\\_app/public/](http://persnds.ufps.edu.co/pers_app/public/)
- [61] IDEAM, "Atlas de Radiación Solar, Ultravioleta y Ozono de Colombia," [Online]. Available: <http://atlas.ideam.gov.co/visorAtlasRadiacion.html>. [Accessed 2018].
- [62] IDEAM, "EVALUACIÓN DE LA IRRADIACIÓN GLOBAL HORIZONTAL EN COLOMBIA," 2018. [Online]. Available: <http://atlas.ideam.gov.co/basefiles/Evaluacion-de-la-Irradiacion-Global-Horizontal-en-Colombia.pdf>. [Accessed 2018 August].
- [63] IDEAM, "Catálogo Nacional de Estaciones del IDEAM," 2018. [Online]. Available: <https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Cat-logo-Nacional-de-Estaciones-del-IDEAM/hp9r-jxuu/data>. [Accessed August 2018].
- [64] IDEAM, "ANEXO: LISTA DE ESTACIONES AUTOMÁTICAS SATELITALES DE RADIACIÓN GLOBAL DEL IDEAM USADAS EN EL ATLAS," 2018. [Online]. Available: <http://atlas.ideam.gov.co/basefiles/Anexo-Lista-de-estaciones-automaticas-de-radiacion-global-del-Ideam.pdf>. [Accessed August 2018].
- [65] IDEAM, "Glosario - IDEAM," 2018. [Online]. Available: <http://www.ideam.gov.co/web/atencion-y-participacion-ciudadana/glosario>. [Accessed August 2018].
- [66] IDEAM, "EVALUACIÓN DEL BRILLO SOLAR EN COLOMBIA," [Online]. Available: <http://atlas.ideam.gov.co/basefiles/Evaluacion-del-brillo-solar-en-Colombia.pdf>. [Accessed August 2018].
- [67] MathWorks, "Análisis de datos," 2018. [Online]. Available: [https://es.mathworks.com/help/matlab/learn\\_matlab/data-analysis.html](https://es.mathworks.com/help/matlab/learn_matlab/data-analysis.html). [Accessed September 2018].
- [68] R. Khoury and D. Wilhelm Harder, "Interpolation, Regression, and Extrapolation," in *Numerical methods and modelling for engineering*, Springer, 2016, pp. 77-110.
- [69] D. Kahaner, C. Moler and S. Nash, "Interpolation," in *Numerical methods and software*, New Jersey, USA, Prentice Hall, 1989, pp. 81-106.
- [70] F. N. Fritsch and R. E. Carlson, "Monotone Piecewise Cubic Interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, p. 238–246, 1980.
- [71] M. Paulescu, E. Paulescu, P. Gravila and V. Badescu, "State of the Sky Assessment - Clearness index," in *Weather Modeling and Forecasting of PV Systems Operation*, London, UK, Springer, 2013, pp. 43-86.

- [72] L. Seungyeoun, O. Jinseok and K. Min-Seok, "Gene-gene interaction analysis for the survival phenotype based on the standardized residuals from parametric regression models," in *IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Atlanta, GA, USA, 2011.
- [73] C. Yau, "Standardized Residual - R Tutorial eBook," r-tutor.com, 2018. [Online]. Available: <http://www.r-tutor.com/elementary-statistics/simple-linear-regression/standardized-residual>. [Accessed September 2018].
- [74] Wikipedia, "Errors and residuals," 14 September 2018. [Online]. Available: [https://en.wikipedia.org/wiki/Errors\\_and\\_residuals](https://en.wikipedia.org/wiki/Errors_and_residuals). [Accessed September 2008].
- [75] Mathworks, "Residuals," 2018. [Online]. Available: <https://es.mathworks.com/help/stats/residuals.html>. [Accessed September 2018].
- [76] Mathworks, "Residual Analysis," 2018. [Online]. Available: <https://es.mathworks.com/help/curvefit/residual-analysis.html>. [Accessed September 2018].
- [77] Mathworks, "Evaluating Goodness of Fit," 2018. [Online]. Available: <https://es.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html>. [Accessed September 2018].
- [78] S. Pop, I. Ciascai and D. Pitica, "Statistical Analysis of Experimental Data Obtained from the Optical Pendulum," in *IEEE 16th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, Pitesti, Romania, 2010.
- [79] F. Russo, S. A. Biancardo and A. F. Gianluca Dell'Acqua, "Predicting percent air voids content in compacted bituminous hot mixture specimens by varying the energy laboratory compaction and the bulk density assessment method," *Construction and Building Materials*, vol. 164, p. 508–524, 2018.
- [80] S. Çınar and N. Acir, "A novel system for automatic removal of ocular artefacts in EEG by using outlier detection methods and independent component analysis," *Expert Systems With Applications*, vol. 68, p. 36–44, 2017.
- [81] J. R. Taylor, "Chapter 6: Rejection data," in *An introduction to error analysis - The study of uncertainties in physical measurements*, Sausalito, California, USA, University Science Books, 1982, pp. 165-173.
- [82] A. DasGupta, "Chapter 9: Normal Distribution," in *Fundamentals of Probability: A first course*, West Lafayette, IN, USA, Springer, 2010, pp. 195-212.
- [83] S. M. Ross, "Chapter 8: Hypothesis Testing," in *Introduction to probability and statistics for engineers and scientists*, San Diego, California, USA, Elsevier Academic Press, 2004, pp. 291-351.
- [84] E. Akarslan and F. O. Hocaoglu, "A novel adaptive approach for hourly solar radiation forecasting," *Renewable Energy*, vol. 87, pp. 628-633, 2016.
- [85] F.-V. Gutierrez-Corea, M.-A. Manso-Callejo, M.-P. Moreno-Regidor and M.-T. Manrique-Sancho, "Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations," *Solar Energy*, vol. 134, pp. 119-131, 2016.
- [86] Mathworks, "Improve Shallow Neural Network Generalization and Avoid Overfitting," 2018. [Online]. Available: <https://es.mathworks.com/help/deeplearning/ug/improve-neural-network-generalization-and-avoid-overfitting.html>. [Accessed October 2018].
- [87] Mathworks, "trainlm," 2018. [Online]. Available: <https://es.mathworks.com/help/deeplearning/ref/trainlm.html;jsessionid=9747022f67c7542bc42a7d6f11ce>. [Accessed October 2018].
- [88] I. Khan, H. Zhu, D. Khan and M. Kumar Panjwani, "Photovoltaic Power prediction by Cascade forward artificial neural network," in *International Conference on Information and Communication Technologies (ICICT)*, Karachi, Pakistan, 2017.
- [89] R. Meenal and A. I. Selvakumar, "Review on artificial neural network based solar radiation prediction," in *2nd International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2017.

- [90] O. Çelik, A. Teke and H. B. Yıldırım, "The optimized artificial neural network model with LevenbergeMarquardt algorithm for global solar radiation estimation in Eastern Mediterranean Region of Turkey," *Journal of Cleaner Production*, vol. 116, pp. 1-12, 2016.
- [91] A. Khosravi, R. Koury, L. Machado and J. Pabon, "Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms," *Journal of Cleaner Production*, vol. 176, pp. 63-75, 2018.
- [92] M. Shaddel, D. S. Javan and P. Baghernia, "Estimation of hourly global solar irradiation on tilted absorbers from horizontal one using Artificial Neural Network for case study of Mashhad," *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 59-67, 2016.
- [93] C. Renno, F. Petito and A. Gatto, "ANN model for predicting the direct normal irradiance and the global radiation for a solar application to a residential building," *Journal of Cleaner Production*, vol. 135, pp. 1298-1316, 2016.
- [94] B. Ihya, A. Mechaqrane, R. Tadili and M. N. Bargach, "Prediction of hourly and daily diffuse solar fraction in the city of Fez (Morocco)," *Theoretical and Applied Climatology - Springer*, vol. 120, pp. 737-749, 2015.
- [95] X. Xue, "Prediction of daily diffuse solar radiation using artificial neural networks," *international journal of hydrogen energy*, vol. 42, pp. 28214-28221, 2017.
- [96] L. Zou, L. Wang, A. Lin, H. Zhu, Y. Peng and Z. Zhao, "Estimation of global solar radiation using an artificial neural network based on an interpolation technique in southeast China," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 146, pp. 110-122, 2016.
- [97] M. Marzouq, Z. Bounoua, H. El Fadili, A. Mechaqrane, K. Zenkouar and Z. Lakhliai, "New daily global solar irradiation estimation model based on automatic selection of input parameters using evolutionary artificial neural networks," *Journal of Cleaner Production*, vol. 209, pp. 1105-1118, 2018 (Online).
- [98] Fraunhofer Institute for Solar Energy Systems, ISE with support of PSE Conferences & Consulting GmbH, "PHOTOVOLTAICS REPORT," 26 February 2018. [Online]. Available: [www.ise.fraunhofer.de](http://www.ise.fraunhofer.de). [Accessed Sept. 2018].
- [99] M. A. Omar and M. M. Mahmoud, "Grid connected PV- home systems in Palestine: A review on technical performance, effects and economic feasibility," *Renewable and Sustainable Energy Reviews*, vol. 82, no. 1, pp. 2490-2497, 2018.
- [100] S. Rodrigues, R. Torabikalaki, F. Faria, N. Cafôfo, X. Chen, A. R. Ivaki, H. Mata-Lima and F. Morgado-Dias, "Economic feasibility analysis of small scale PV systems in different countries," *Solar Energy*, vol. 131, no. 1, p. 81-95, 2016.
- [101] SunFields – Solar panels and inverter Supplier, "¿Cuál es el mejor panel solar fotovoltaico del mundo?," 2018, [Online]. Available: <https://www.sfe-solar.com/paneles-solares/el-mejor-panel-solar-del-mundo/>. [Accessed Sept.].
- [102] TÜV Rheinland, "PV+Test 2.0 from TÜV Rheinland and Solarpraxis: Benchmark test for photovoltaic modules expanded," [Online]. Available: [https://www.tuv.com/media/japan/newsevent/solar\\_tech\\_news/2013\\_no\\_1/english\\_4/STN1\\_PVTest20\\_E.pdf](https://www.tuv.com/media/japan/newsevent/solar_tech_news/2013_no_1/english_4/STN1_PVTest20_E.pdf). [Accessed Sept. 2018].
- [103] A. Sagani, J. Mihelis and V. Dedoussis, "Techno-economic analysis and life-cycle environmental impacts of small-scale building-integrated PV systems in Greece," *Energy and Buildings*, vol. 139, no. 1, p. 277-290, 2017.
- [104] D. A. Quansah, M. S. Adaramola, G. K. Appiah and I. A. Edwin, "Performance analysis of different grid-connected solar photovoltaic (PV) system technologies with combined capacity of 20 kW located in humid tropical climate," *International journal of hydrogen energy*, vol. 42, pp. 4626-4635, 2017.
- [105] E. Banguero, A. J. Aristizábal and W. Murillo, "A Verification Study for Grid-Connected 20 kW Solar PV System Operating in Chocó, Colombia," *Energy Procedia*, vol. 141, pp. 96-101, 2017.
- [106] J. Lucio, R. Valdés and L. Rodríguez, "Loss-of-load probability model for stand-alone photovoltaic systems in Europe," *Solar Energy*, vol. 86, p. 2515-2535, 2012.

- [107] M. Egido and E. Lorenzo, "The sizing of stand alone PV-systems: a review and a proposed new method," *Solar Energy Materials and Solar Cells*, vol. 26, pp. 51-69, 1992.
- [108] K. Benmouiza, M. Tadj and A. Cheknane, "Classification of hourly solar radiation using fuzzy c-means algorithm for optimal stand-alone PV system sizing," *Electrical Power and Energy Systems*, vol. 82, p. 233–241, 2016.
- [109] A. Mellit, "ANN-based GA for generating the sizing curve of stand-alone photovoltaic systems," *Advances in Engineering Software*, vol. 41, p. 687–693, 2010.
- [110] R. Posadillo and R. López Luque, "Approaches for developing a sizing method for stand-alone PV systems with variable demand," *Renewable Energy*, vol. 33, pp. 1037-1048, 2008.
- [111] A. I. Ibrahim, T. Khatib and A. Mohamed, "Optimal sizing of a standalone photovoltaic system for remote housing electrification using numerical algorithm and improved system models," *Energy*, vol. 126, pp. 392-403, 2017.
- [112] A. W. Grace and I. A. Wood, "Approximating the tail of the Anderson–Darling distribution," *Computational Statistics and Data Analysis*, vol. 56, p. 4301–4311, 2012.
- [113] P. Closas, J. Arribas and C. Fernández-Prades, "Testing for normality of UWB-based distance measurements by the Anderson-Darling statistic," in *Future Network & Mobile Summit - IEEE*, Florence, Italy, 2010.
- [114] D. Han, X. Tan and P. Shi, "Clutter Distribution Identification Based on Anderson-Darling Test," in *IEEE International Conference on Computer and Communications*, Chengdu, China, 2017.
- [115] A. H-S. Ang and W. H. Tang, "Chapter 7.3: Testing goodness of fit of distribution models," in *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering*, New York, USA, Wiley, 2007.
- [116] S. M. Ross, "Chapter 5: Special Random Variables," in *Introduction to probability and statistics for engineers and scientists*, San Diego, California, USA, Elsevier Academic Press, 2004, pp. 185-192.
- [117] Unidad de Planeación Minero Energética (UPME) - IDEAM, "Atlas de Radiación Solar en Colombia," 2005. [Online]. Available: <https://biblioteca.minminas.gov.co/pdf/Atlas%20de%20radiaci%C3%B3n%20solar%20Colombia.pdf>. [Accessed September 2018].
- [118] J. W. Spenser, "Fourier Series Representation of the Position of the Sun," *Search*, vol. 2, no. 5, p. 172, 1971.
- [119] Y. Jiang, "Calculation of daily global solar radiation for Guangzhou, China," in *International Conference on Optics, Photonics and Energy Engineering*, Wuhan, China, 2010.
- [120] A. Nasser Eddine and I. Hage Chehade, "Estimation model for global solar radiation in Lebanon," in *3rd International Conference on Renewable Energies for Developing Countries (REDEC)*, Zouk Mosbeh, Lebanon, 2016.