sustainability

MDPI

Article

Monthly Global Solar Radiation Model Based on Artificial Neural Network, Temperature Data and Geographical and Topographical Parameters: A Case Study in Spain

Enrique González-Plaza [1], David García [2] and Jesús-Ignacio Prieto [1,*]

[1] Department of Physics, University of Oviedo, c/Federico García Lorca, nº18, 33007 Oviedo, Spain; gonzalezenrique@uniovi.es
[2] Department of Energy, University of Oviedo, c/Wifredo Ricart, s/n, 33204 Gijón, Spain; garciamdavid@uniovi.es
* Correspondence: jprieto@uniovi.es

Abstract: Solar energy plays an essential role in the current energy context to achieve sustainable development while supplying energy needs, creating jobs, and protecting the environment. Many solar radiation models have provided valid estimates at many different locations, using appropriate input variables for specific climatic conditions, but predictions are less accurate on a regional scale. Since radiometric weather stations are relatively dispersed, even in the most developed countries, it is interesting to develop indirect models based on measurements that are common in secondary network stations. This paper develops a monthly global solar radiation model based on a simple neural network structure, using temperature, geographical, and topographical data from 105 meteorological stations, representative of the whole of peninsular Spain. A hierarchical clustering procedure was employed to select the data used to train and validate the model. To avoid functional dependencies between parameters and variables, which hinder the generality of the model, all input and output variables are dimensionless. The estimates fit the 1260 monthly data with RRMSE values of about 6%, which improves results obtained previously, using regression models, and proves that simplicity is compatible with the generality and accuracy of a model, even in large regions with very varied characteristics.

Keywords: global solar radiation; general models; temperature-based models; artificial neural networks; dimensionless variables; hierarchical clustering; Spain

check for updates

Citation: González-Plaza, E.; García, D.; Prieto, J.-I. Monthly Global Solar Radiation Model Based on Artificial Neural Network, Temperature Data and Geographical and Topographical Parameters: A Case Study in Spain. *Sustainability* **2024**, *16*, 1293. https://doi.org/10.3390/su16031293

Academic Editor: Sean Loughney

Received: 14 December 2023
Revised: 18 January 2024
Accepted: 30 January 2024
Published: 2 February 2024

1. Introduction

At the time of writing, the United Nations Climate Change Summit (COP 28) is taking place, from which increasingly urgent global agreements are expected to mitigate the decades-long conflict between social development and respect for the environment. In this context, it is recognised that it is necessary to improve energy efficiency and reduce energy demand as much as possible, accelerating decarbonisation through the increasing use of renewable energies [1–3]. Solar energy is thus gaining in importance even in regions where it has hitherto been considered a resource with insufficient potential [4,5].

Thus, in industrialised countries as well as in less favoured regions, sustainable social development can definitely depend on the existence of reliable information about the availability of global solar radiation (GSR). Such data are generally needed on a local scale, as orographic and microclimatic diversities are particularly influential in project planning [6,7].

Numerous models have been developed over recent decades that provide acceptable indirect estimates of GSR on a horizontal surface using various methodologies and input variables [8]. According to the mathematical methodology, they could be classified as parametric, such as the classical Ångström–Prescott (A-P) model [9,10], and non-parametric, such as those based on artificial neural network (ANN) methods [11]. Depending on the input variables, the models are generally classified into sunshine-based, temperature-based, cloud-based, and other climatic parameters-based models [12,13]. Meteorological satellites are an alternative source of data, particularly interesting in sparsely populated areas. The prediction accuracy depends on the quality of the experimental data, the complexity of the model, i.e., the number and type of influential variables, and the functional relationships. Using site-calibrated coefficients, most models have high accuracy, which generally increases with the functional complexity of the models, but none outperforms the rest anywhere. The use of dimensionally non-homogeneous equations in many recently compiled models can lead to unexpected variations of locally calibrated coefficients, even in similar climates, as these coefficients necessarily depend on variables that are not explicit in the model [13]. The accuracy of a model does not necessarily imply complexity. Simple models, using coefficients calculated from general equations valid for large geographical areas, may be advisable [14]. For sample, a variant of the classical Ångström–Prescott model can explain roughly 90% of the variability in the data recorded at 59 European stations, using relative sunshine, elevation and the month index as input variables [15].

Ground stations with long-term time GSR series are necessary to validate both GSR models and algorithms used for satellite data processing, but the density of radiometric networks is very low worldwide. For example, in China, GSR data are recorded at just over 15% of meteorological stations [16], while sunshine hours data are available at all stations in the Iranian province of Kurdistan, although none have GSR records [17]. In contrast, temperature is measured at all weather stations, since it requires low-cost equipment, but data on sunshine duration or cloudiness are less available. The ratio of radiometric to thermometric stations is no greater than 1:100 in the United States and might be as low as 1:500 globally [16,18]. In short, temperature-based GSR models can facilitate the indirect estimation of solar resources in areas remote from radiometric stations.

Among the most widely used temperature-based models is the one introduced by Hargreaves and Samani (H-S) [19] and Richardson [20], which is based on the relationship between the monthly average of the atmospheric clearness index and the difference between the monthly average daily maximum temperature and the monthly average daily minimum temperature, $\Delta T = T_{\max} - T_{\min}$. This classical model was modified to obtain a dimensionally homogeneous equation, after identifying that the elevation z, the distance to the sea L, and a reference temperature T_{ref}, are influential variables that are implicit in the coefficients of the original model [21]. Using regression techniques, such an approach led to satisfactory results at various scales in coastal areas of northern Spain, where it was compared with other empirical temperature-based models and with meteorological satellite-based databases [14,21–23]. This homogeneous model was used to identify climatic zones in regions with varied orography and proximity to the sea [24–26], and also compared with other temperature-based models in two large areas of Spain, with very different climatology and latitude from those of the northern coast [23]. It was noted that other models may be more accurate at some locations using locally calibrated coefficients; however, the new model obtained the best results when the coefficients of each model were calculated from general equations applicable to the set of stations in each area.

In the present work, a model based on an artificial neural network with a simple structure is developed to estimate the monthly clearness index, using the same variables of the aforementioned homogeneous model with the addition of latitude, and a monthly index, in anticipation of possible seasonal effects. As a case study, the model performance is analysed using data from the Spanish areas previously studied [23]. The results obtained show that the model exceeds the accuracy of previous ones, both for statistical indicators averaged for the set of more than one hundred Spanish stations and for deviations observed in particular stations, which provides general interest to the procedure.

2. Materials and Methods

2.1. Meteorological Stations and Experimental Data

In this work, climate data from 105 stations belonging to official Spanish meteorological networks were considered. The geographical and topographical data of the stations, and the Köppen–Geiger climate classification of their location are provided in Table A1 of Appendix A in Supplementary Data.

In the southern region of Andalusia 71 stations are located, with latitudes between 36.29 °N and 38.30 °N and longitudes between −1.77° E and −7.25° E, from the Mediterranean coast to the Atlantic coast, with elevations between 4.5 m and 1214 m, and distance to the sea between 0.03 km and 184.94 km. Another 13 stations are located in a central area with latitudes between 38.99° N and 42.59° N and longitudes between −1.16° E and −6.34 °E, from the Mediterranean coast to the border with Portugal, with elevations between 53 m and 1085 m, and distance to the sea between 6.43 km and 322.07 km. Finally, 21 stations are located in the northern coastal provinces of Spain, with latitudes between 41.91° N and 43.58° N and longitudes between 2.76° E and −8.42° E, from the Mediterranean coast to the Atlantic coast, with elevations between 12 m and 513 m and distance to the sea between 0.29 km and 55.30 km. According to the Köppen–Geiger climate classification, 61 stations are in the Csa zone, 18 stations in the Cfb zone, 12 stations in the BSk zone, 8 stations in the BSh zone, 5 stations in the Csb zone and a single station in the BWk zone. The variety of geographical, topographical and climatic characteristics can be considered representative of the whole of peninsular Spain. At virtually all stations, records averaged over a minimum of 14 years of the last two decades are used.

For each station, monthly averages over the period of measurements are given as Table A2 of Appendix A in Supplementary Data for the following variables: daily global irradiation on a horizontal surface, daily maximum air temperature and daily minimum air temperature, totalling 3780 data.

2.2. GSR Modelling

Any dimensionally homogeneous physical equation can be rewritten in terms of a set of non-dimensional groups. When the equation cannot be expressed by means of non-dimensional groups, it is said to be incomplete, which is due to the lack of some influencing variable [27]. This procedure is applicable to parametric model equations, but also to transfer functions used in ANN-based models [13]. There are numerous temperature-based models that are not homogeneous, such as the H-S model, represented by the following equation:

$$\frac{H}{H_0} = a_1 \Delta T^{0.5} \tag{1}$$

where the coefficient a_1 was initially set at 0.17 for arid and semiarid climates, but later values of 0.16 and 0.19 were recommended, respectively, for inland and near the coast [28].

Using both the explicit variables in the H-S model and the implicit variables noted by the authors themselves, it was deduced that a complete equation must satisfy the following functional relationship [21]:

$$H = f(H_0, \Delta T, T_{\text{ref}}, z, L), \tag{2}$$

where H is the monthly average of the daily GSR over a horizontal surface, H_0 is the monthly average of the daily extraterrestrial irradiation, and T_{ref} is a reference temperature required by the dimensional homogeneity.

Taking into account Buckingham's theorem and accepting the functional form of the original H-S model, the following model was proposed as a particular case of Equation (2):

$$\frac{H}{H_0} = a_1 \left(\frac{\Delta T}{T_{\text{ref}}} \right)^{0.5} \tag{3}$$

where $a_1 = f(z/L)$ is a characteristic coefficient of each locality.

Equation (3) was used in previous GSR analyses using locally calculated coefficients and identifying the reference temperature as T_{min} from experimental results. It was also deduced that a_1 can be calculated by the following general equation at locations with not very different latitudes:

$$a_1 = a - b\,\mathrm{e}^{-cz/L}, \tag{4}$$

where a, b and c are characteristic parameters of the region under consideration [21,23].

As mentioned above, the procedure based on Equations (3) and (4) provided acceptable results in each of the three broad Spanish areas previously studied [23]. To extend the model to all three areas, it is necessary to include latitude ϕ in the list of influential variables, as well as an index of the mean monthly day d associated with each datum, in anticipation of possible seasonal influences. Therefore, the extended model is based on the following functional relationship:

$$H = f(H_0, \Delta T, T_{\mathrm{min}}, z, L, \phi, d), \tag{5}$$

from which Buckingham's theorem gives the following equivalent expression:

$$\frac{H}{H_0} = f\left(\frac{\Delta T}{T_{\mathrm{ref}}}, \frac{z}{L}, \phi, d\right). \tag{6}$$

This functional relationship is the basis of the ANN developed in this work, with $\Delta T/T_{\mathrm{ref}}$, z/L, ϕ and d as input variables and H/H_0 as output variable. The final structure of the network is determined using data from representative stations in each area for training various configurations. For each configuration, the evaluation of results is based on optimising the statistical indicators averaged for the set of training stations. Finally, the averaged statistical indicators are evaluated for the total of training and test stations, as well as the deviations obtained for monthly values at certain stations.

2.3. Hierarchical Clustering of Data

Clustering techniques allow the selection of representative stations to be used for neural network training, so that acceptable errors are obtained at each station during the validation process.

Among the different clustering algorithms available in the literature, a bottom-up hierarchical clustering was selected to classify the stations [29,30]. The clustering process was performed assuming that the representativeness of a weather station is independent of the seasonality, i.e., disregarding the influence of the monthly day index, d. Consequently, based on Equation (6), a Euclidean metric in three-dimensional space was defined to calculate the distance between stations, using annual averages of $\Delta T/T_{\mathrm{ref}}$, z/L and ϕ as the coordinates. Due to the huge differences between the ranges of variation of these coordinates, the variables were normalised to the interval [0, 1] using the following equations:

$$\left(\frac{\Delta T}{T_{\mathrm{ref}}}\right)_i^* = \frac{\left(\frac{\Delta T}{T_{\mathrm{ref}}}\right)_i - \min\limits_{j=1\ldots n}\left\{\left(\frac{\Delta T}{T_{\mathrm{ref}}}\right)_j\right\}}{\max\limits_{j=1\ldots n}\left\{\left(\frac{\Delta T}{T_{\mathrm{ref}}}\right)_j\right\} - \min\limits_{j=1\ldots n}\left\{\left(\frac{\Delta T}{T_{\mathrm{ref}}}\right)_j\right\}} \tag{7}$$

$$\left(\frac{z}{L}\right)_i^* = \frac{\left(\frac{z}{L}\right)_i - \min\limits_{j=1\ldots n}\left\{\left(\frac{z}{L}\right)_j\right\}}{\max\limits_{j=1\ldots n}\left\{\left(\frac{z}{L}\right)_j\right\} - \min\limits_{j=1\ldots n}\left\{\left(\frac{z}{L}\right)_j\right\}} \tag{8}$$

$$\phi_i^* = \frac{\phi_i - \min\limits_{j=1\ldots n}\{\phi_j\}}{\max\limits_{j=1\ldots n}\{\phi_j\} - \min\limits_{j=1\ldots n}\{\phi_j\}} \tag{9}$$

In these equations, n is the number of stations, and $(\Delta T/T_{\text{ref}})_i$ is the annual average value for i-th station, which is obtained from the monthly average values $(\Delta T/T_{\text{ref}})_i^m$ by means of the following equation:

$$\left(\frac{\Delta T}{T_{\text{ref}}}\right)_i = \frac{1}{12}\sum_{m=1}^{12}\left(\frac{\Delta T}{T_{\text{ref}}}\right)_i^m, \tag{10}$$

Thus, the normalised distance d_{ij} between stations was computed using the usual Euclidian metric, i.e.:

$$d_{ij} = \sqrt{\left(\left(\frac{\Delta T}{T_{\text{ref}}}\right)_i^* - \left(\frac{\Delta T}{T_{\text{ref}}}\right)_j^*\right)^2 + \left(\left(\frac{z}{L}\right)_i^* - \left(\frac{z}{L}\right)_j^*\right)^2 + \left(\phi_i^* - \phi_j^*\right)^2}. \tag{11}$$

2.4. ANN Characteristics

According to its biological definition, a neuron is a living cell placed within a network of them which is given inputs or excitations and finally yields an output or interpretation based on internal processes [31]. Regarding the case of an ANN, a neuron can be classified into three groups, as shown in Figure 1. The input layer acts as the entrance of the numerical data of the variables of interest $x_1, \ldots, x_n$ used to estimate the values $a_1, \ldots, a_k$ of the network, and provides information to the neurons located within the hidden layers, whose inputs and outputs are not accessible from the outside. A neural network can present as many hidden layers as necessary to achieve the expected behaviour, although the complexity of its structure increases with the number of neurons and hidden layers and thus, the time needed for the optimisation process.

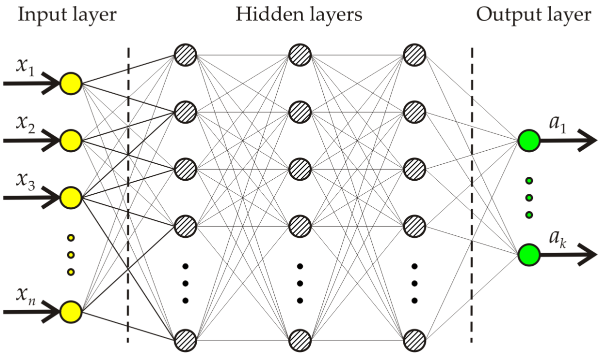


Figure 1. Architecture of an artificial neural network.

The neuron output is obtained by applying Equation (12). The parameter y_j depicts the output value of the j-th neuron, which is calculated as the outcome of a certain activation function f, given a linear combination of the output values of the previous layer. Thus, x_n is the input of the n-th neuron of the layer, w_{nj} is the weight or gain factor of each excitation from the previous layer, and k_j is the independent term. According to the literature, the activation function could be, among other options, an arctangent, a ramp, or a sigmoid function [32], being the ramp the one used in this work since the values of solar radiation can only be positive. The foundation behind its usage is to introduce a non-linear behaviour in the system.

$$y_j = f\left(\sum_{i=1}^{n} w_{ij}x_i + k_j\right) \tag{12}$$

Consequently, the nodes of the output layer provide the calculation of the whole ANN using the outputs of the outermost hidden layer. These final values are computed in the same way as in any other node.

Considering the structure of an ANN, the number of variables involved in the network is big enough to make it unfeasible to find directly the best configuration. Therefore, an optimisation or training process is required to reach the final one. During this process, it is necessary to define a certain cost function and to use a convergence algorithm, whose objective is to obtain for each neuron the weights and biases defined in Equation (12), that optimise the function. The overall behaviour of the results should be evaluated by comparing the ANN output with the objective values using a statistical indicator.

There are several algorithms that can be used during the process, such as genetics or gradient variants. In general, no algorithm could be regarded as explicitly better than the rest, although, given the characteristics of an ANN, some may achieve the solution quicker than others, or be more suitable for more complex configurations. In this sense, genetic algorithms may be more adequate for the latter since they are likelier to avoid getting stuck in a local optimum of the cost function. Nonetheless, the scope of this work does not lie within the study of the best optimisation algorithm, but to analyse the feasibility of a unique solar radiation model for a certain climatic diverse territory, like Spain, using an ANN. Therefore, a gradient algorithm is utilized to find the final parameters of the ANN during the training process. In this process, a step variable is defined and then, in each iteration, this value is summed or subtracted from a random weight or independent term. If the cost function is improved, the modification is kept. Furthermore, to avoid a possible overfitting of the ANN, the number of optimal neurons and hidden layers are pending further adequacy analysis, based on the results obtained for the test data. The stopping criterion was improving the performance of the previous regression models or reaching a certain number of iterations based on the complexity of the ANN.

2.5. Statistical Indicators

Many statistical indicators have been used in the literature to assess the fit between measurements and model predictions [33]. Since the percentage results facilitate interpretations, in this article the performance of the models is assessed using dimensionless statistical indicators, namely the relative root mean square error RRMSE, the relative mean bias error RMBE, and the coefficient of determination R^2. To facilitate comparisons with results from other authors, the normalised values of the root mean square error and mean bias error, NRMSE and NMBE, have also been calculated, using the mean values of the experimental data as references.

However, none of the statistical indicators provides complete information by itself [34], thus, in this work the values of RRMSE and R^2 are used as main indicators for model comparison. The information is completed with the normalised centred pattern root mean square error, E'_n, calculated as follows:

$$E'_n = \sqrt{1 + \sigma_{sn}^2 - 2\sigma_{sn} R}. \tag{13}$$

This equation and the law of cosines are the basis for a variant of the Taylor diagram [35] using dimensionless variables, which allows the performance of different models to be compared graphically at once and with maximum generality. As can be seen in Figure 2, the performance of a model increases with its proximity to the reference, for which the model estimates would match the experimental data, i.e., $\sigma_{sn} = R = 1$ and $E'_n = 0$.

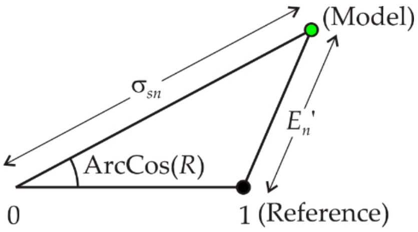


Figure 2. Geometrical basis of the Taylor diagram using normalised variables.

Following the methodology of previous works [13,14,21,23], the RRMSE value was used as the cost function in the optimising algorithm. Nevertheless, the assessment of the goodness-of-fit of the model for the set of stations using the above-mentioned statistical indicators is complemented by a local error analysis at each station using monthly data.

3. Results

3.1. Number of Clusters and Selection of Training and Testing Data

At the beginning of the bottom-up hierarchical clustering procedure, each station forms a cluster by itself and the normalised distance between each station and the others is calculated. The two closest stations are merged into a new cluster, which is assumed to have the mean coordinates of the two original stations, and the sequence is repeated until all stations are merged into a single cluster.

The graphical representation of the clustering procedure is the dendrogram shown in Figure 3. The best choice of the number of clusters is the number of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster [29]. Therefore, the dendrogram leads to four clusters, composed of the stations listed in Table 1. Figure 4 shows the representation of the stations in clusters using normalised coordinates in the 3D space.

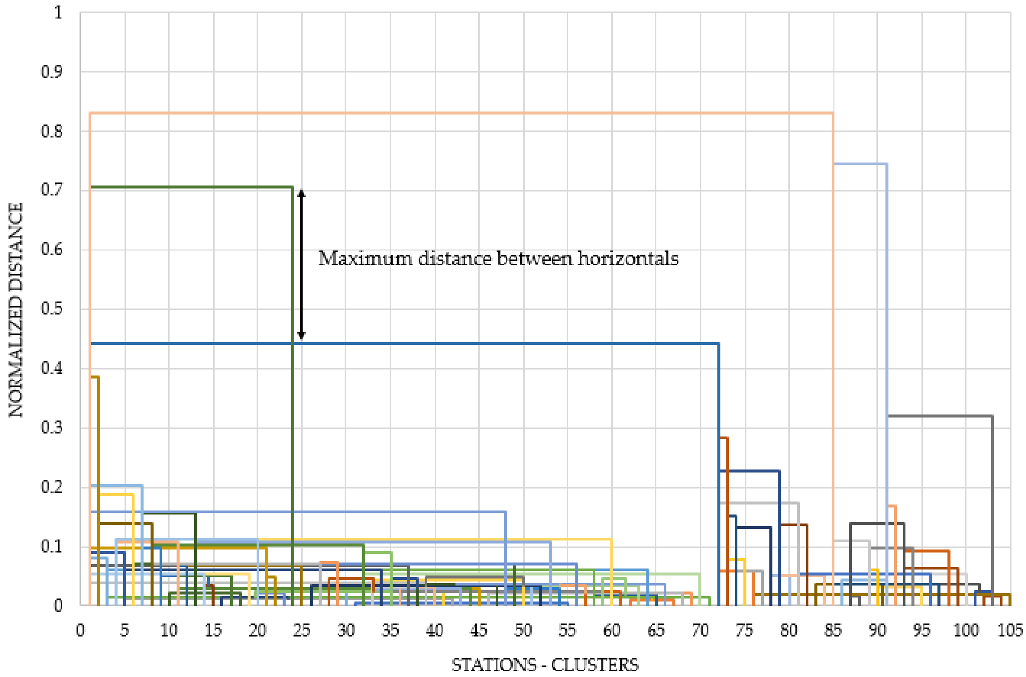


Figure 3. Dendrogram derived from bottom-up hierarchical clustering.

Table 1. Distribution of stations in each cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
91, 92, 103	24	85, 86, 87, 88, 89, 90, 93, 94, 95, 97, 98, 99, 100	Rest of stations

Half of the stations in each cluster were randomly selected for training or validation, amounting to a total of 53 training stations and 52 testing stations. Table 2 lists the distribution of the training stations in the clusters.

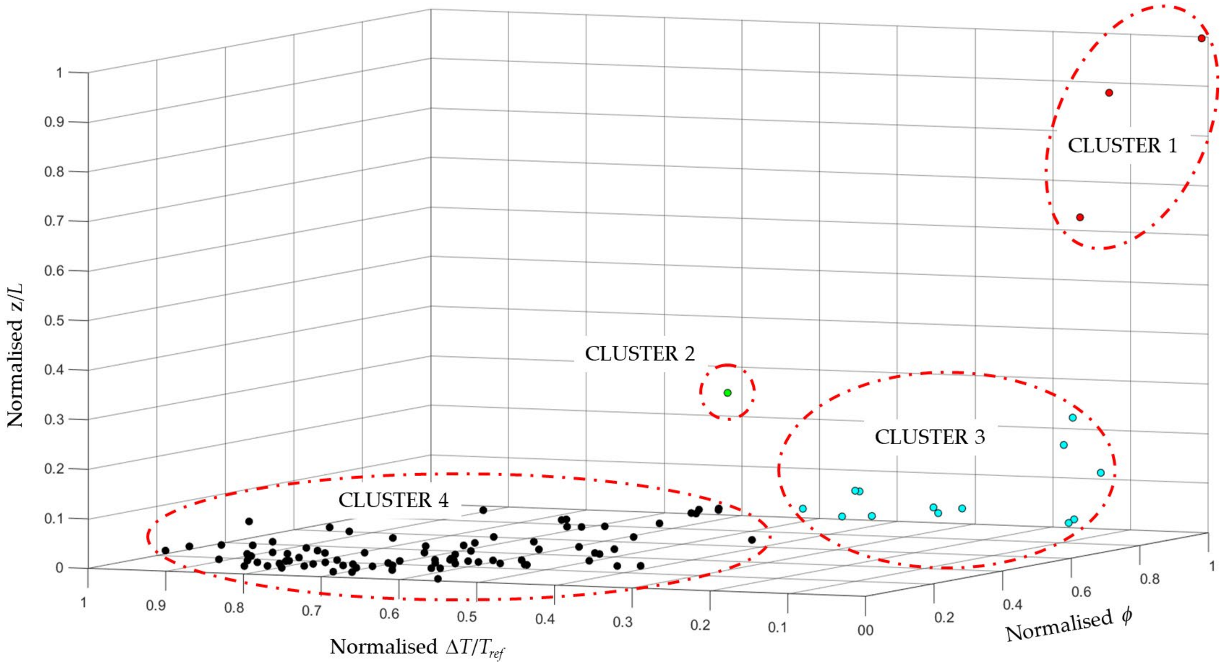


Figure 4. Distribution of stations in normalised space.

Table 2. Distribution of training stations in each cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
91, 103	24	86, 88, 90, 93, 95, 97, 99	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 101, 105

3.2. Selection of the ANN Architecture

Different configurations were tested ranging from one to three hidden layers and having between 1 and 15 neurons in each one of them. The best RRMSE values obtained for the training stations correspond to an ANN with 3 hidden layers and 15 neurons in each layer, so this configuration was adopted, pending the results for the testing stations.

Figure 5a provides a graphical comparison between RRMSE and R2 values obtained with all configurations analysed, while Figure 5b provides an immediate comparison between configurations using three hidden layers.

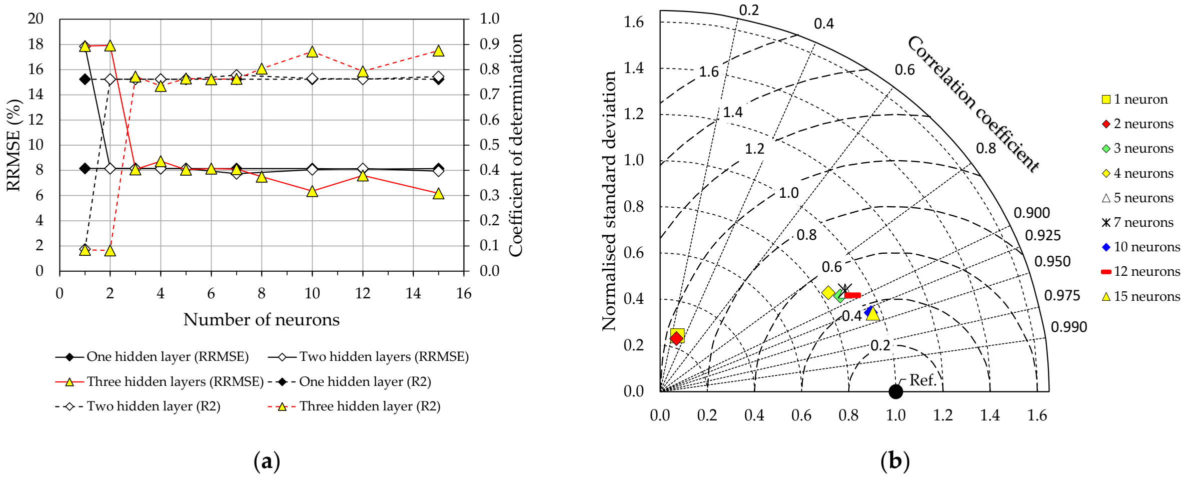


Figure 5. Comparison between ANN configurations: (**a**) Variation of RRMSE and R2 with the number of neurons and hidden layers; (**b**) Taylor diagram for three hidden layers.

Figure 6 shows the comparison between the experimental monthly data and the model predictions. As can be seen, most of the results are within the dashed lines of ±15% relative error, which is in agreement with the RRMSE value close to 6% obtained for the total of 636 monthly data, as indicated in Table 3.

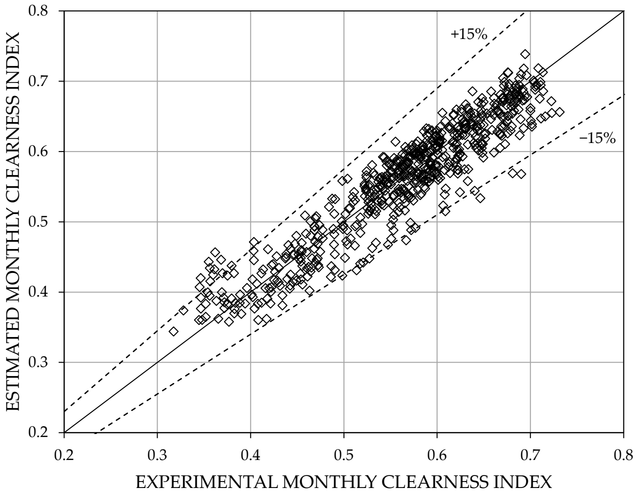


Figure 6. Comparison between experimental data and ANN estimates for training stations.

Table 3. Summary of statistical indicators for several ANN configurations.

	Neurons	**1**	**2**	**3**	**4**	**5**	**7**	**10**	**12**	**15**
One hidden layer	RRMSE (%)	8.15	8.15	8.15	8.15	8.15	8.15	8.15	8.10	8.15
	NRMSE (%)	7.97	7.98	7.97	7.97	7.97	7.97	7.97	7.94	7.97
	RMBE (%)	−0.66	−0.66	−0.66	−0.66	−0.66	−0.66	−0.66	−0.66	−0.66
	NMBE (%)	−1.25	−1.25	−1.25	−1.25	−1.24	−1.25	−1.25	−1.23	−1.25
	R^2	0.7621	0.7616	0.7624	0.7621	0.7621	0.7621	0.7621	0.7635	0.7621
	σ_{sn}	0.898	0.897	0.898	0.898	0.898	0.898	0.898	0.898	0.898
	E_n'(%)	48.83	48.88	48.83	48.83	48.83	48.83	48.83	48.69	48.83
Two hidden layers	RRMSE (%)	17.84	8.15	8.15	8.15	8.15	7.74	8.06	8.15	7.94
	NRMSE (%)	16.48	7.97	7.97	7.97	7.97	7.67	7.92	7.97	7.79
	RMBE (%)	−3.18	−0.66	−0.66	−0.66	−0.66	−0.60	−0.65	−0.66	−0.63
	NMBE (%)	−5.84	−1.24	−1.24	−1.24	−1.24	−1.13	−1.22	−1.24	−1.18
	R^2	0.0869	0.7621	0.7621	0.7621	0.7621	0.7786	0.7650	0.7621	0.7725
	σ_{sn}	0.260	0.898	0.898	0.898	0.898	0.909	0.898	0.898	0.904
	E_n'(%)	95.62	48.84	48.84	48.84	48.84	47.12	48.53	48.84	47.78
Three hidden layers	RRMSE (%)	17.87	17.93	8.08	8.73	8.05	8.10	6.36	7.60	6.18
	NRMSE (%)	16.51	16.54	7.83	8.50	7.91	7.94	5.82	7.42	5.74
	RMBE (%)	−3.19	−3.21	−0.82	−1.00	−0.65	−0.66	−0.41	−0.58	−0.38
	NMBE (%)	−5.85	−5.89	−1.50	−1.82	−1.22	−1.24	−0.73	−1.07	−0.68
	R^2	0.0820	0.082	0.7726	0.7350	0.7653	0.7635	0.8720	0.7934	0.8758
	σ_{sn}	0.253	0.240	0.869	0.832	0.898	0.898	0.957	0.918	0.965
	E_n'(%)	95.77	95.92	47.70	51.54	48.50	48.69	35.85	45.57	35.36

For further validation, ANN performance was studied with data from test stations, with similar results to those obtained using training data, as shown in Table 4. Figure 7 provides the graphical comparison between ANN estimates and experimental data, showing that, for both training and testing data, most monthly results have a relative error of less than ±15%, which is consistent with the RRMSE value of close to 6%. During the training process, no regularization technique was used. The reason for this decision was to use the same cost function as in the case of the linear regression models. Since the results of the testing process are similar to the training results, it is observed that the decision did not lead to an apparent over-fitting.

Table 4. Statistical indicators of the performance of the training and testing data.

	RRMSE (%)	**NRMSE (%)**	**RMBE (%)**	**NMBE (%)**	R^2	σ_{sn}	$E_n^{'}$ **(%)**
Training	6.18	5.74	−0.38	−0.68	0.8758	0.9646	35.36
Testing	6.63	6.10	−0.25	−0.48	0.8519	0.9950	39.16

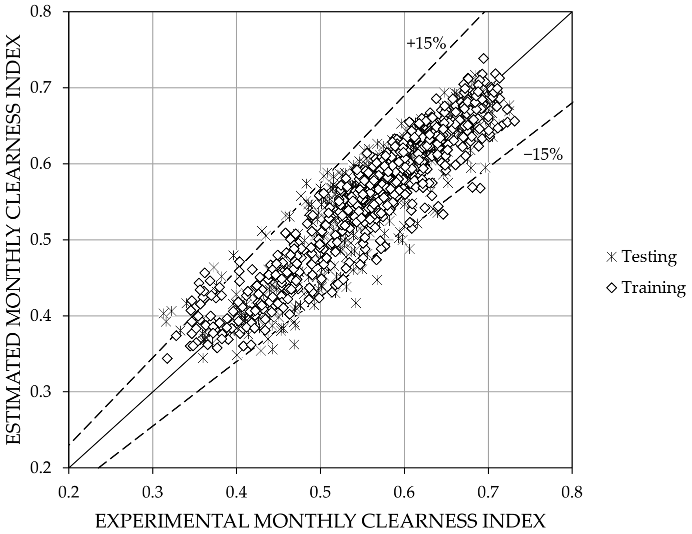


Figure 7. Comparison between experimental data and ANN estimates for training and testing stations.

4. Discussion

The hierarchical clustering procedure, based on annual averages of the relative temperature difference, the ratio between elevation and distance to the sea, and latitude, led to four clusters to classify the 105 stations located in three areas of peninsular Spain. Cluster 1 contains stations No. 91, 92 and 103, all of them located close to the sea on the northern coast of Spain, with the highest z/L values. Cluster 2 consists only of station No. 24, which also has a high z/L value. These clusters, formed by a few stations, present peculiarities, either from a climatic, geographical or topographical point of view.

The model of Equation (3) was compared for the first time with 13 other temperature-based models of low or moderate functional complexity [23], using the same stations and meteorological data of the present article. The coefficients of each model were calculated as a function of the z/L ratio using equations obtained by regression techniques from the total data in each area. Table 5 shows in bold characters the best statistical averages calculated by this procedure for the set of stations in each area. Depending on the area, these results were obtained by means of Equations (1) and (3), or the model proposed by Adaramola for various locations in Nigeria [36], which is expressed by the following equation:

$$\frac{H}{H_0} = a_1 + a_2 T_{\rm m} \tag{14}$$

Table 5. Comparison between best-performing models in each area using coefficients derived from regression equations [23].

	Stations No. 1–71		Stations No. 72–84		Stations No. 85–105	
Model	**RRMSE (%)**	E_n'**(%)**	**RRMSE (%)**	E_n'**(%)**	**RRMSE (%)**	E_n'**(%)**
Equation (1)	9.10	23.36	**7.00**	**14.31**	9.94	23.50
Equation (3)	9.54	24.23	7.29	15.16	**8.91**	**20.15**
Equation (14)	**4.51**	**10.08**	11.18	23.42	12.41	29.52

With the data and procedure used, there is no doubt that the model of Equation (14) outperforms other regression models for the 71 stations in Andalusia, while the models based on Equations (1) and (3) are more appropriate for the other two regions. In this comparison, it should be noted that each equation has numerically different coefficients in each of the three regions, which were derived by calibration using data from all stations in each region. Averaging the model errors of Equations (1), (3) and (14) for the total of

105 stations leads to RRMSE values of 9.05%, 9.16% and 7.75%, respectively. Therefore, the ANN-based model shows acceptable accuracy for the total of 105 stations, with an RRMSE value close to 6% and moderate local monthly GSR errors, as deduced from Figures 6 and 7, and Table 4. It should also be noted that the results of Adaramola's model are favoured in Andalusia because the z/L ratio is high only at station No. 24, where this model obtains an annual averaged value of RRMSE = 50.32% and is therefore not acceptable from the perspective of local performance [23]. This previous observation is consistent with the fact that station No. 24 is not included in the same cluster as other stations in Andalusia, and shows that averages calculated with station distributions in different z/L ratio ranges can mask unacceptable local results.

Regarding the ANN performance at the local scale, Table 6 and Figure 8 show the number of monthly GSR estimates with relative error less than a given value, and the percentage of the total 1260 training and test data. It follows that almost 90% of local monthly values have a relative error of less than 10% and that no value exceeds 30%. It can be observed in Tables B1 and B2 of Appendix B in Supplementary Data that most local monthly GSR estimates have relative errors within ±5%, without remarkable differences between training and testing values. The monthly relative errors are above 15% for 7 training stations and 11 testing stations, which are plotted in Figure 9. As can be seen, these errors cannot be associated with certain ranges of the z/L ratio or the latitude ϕ, since these variables fall, respectively, in the intervals $2.97 \leq z/L \leq 239.09$ and $40.959° \leq \phi \leq 43.531°$ for the training stations, and $1.90 \leq z/L \leq 333.3$ and $37.060° \leq \phi \leq 43.584°$ for the testing stations.

Table 6. Number of monthly GSR estimates with relative error less than a given value and percentage relative to total training and testing data.

x	3.0%	6.0%	10.0%	12.5%	15.0%	17.5%	20.0%	22.5%	27.5%	30.0%
Number of values	515	905	1108	1162	1202	1236	1246	1252	1259	1260
P (RE < x)	40.9%	71.8%	87.9%	92.2%	95.4%	98.1%	98.9%	99.4%	99.9%	100.0%

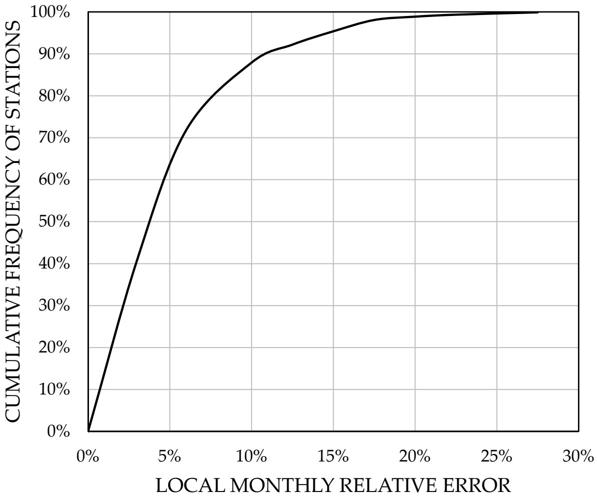
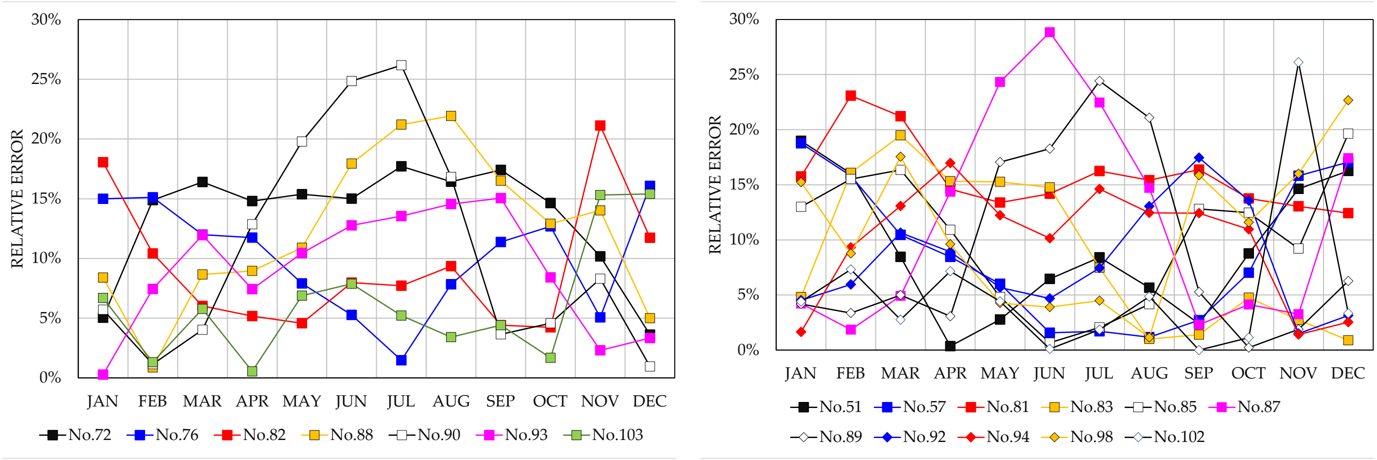


Figure 8. Cumulative frequency of stations versus local monthly relative errors.

Therefore, the ANN-based model turns out to be more accurate than those based on regression techniques, because, with the same variables, it not only provides acceptable statistical averages for the three areas studied as a whole, but also lower monthly relative errors at the local scale. Since the inclusion of the parameter z/L and latitude ϕ makes it possible to account for local characteristics, the proposed model acquires generality. As future work, it is planned to immediately evaluate the extension of the model to the whole

Iberian Peninsula and, at a later stage, to other regions, as well as research on the inclusion of additional input variables to reduce the monthly relative errors.



(**a**) (**b**)

Figure 9. Highest monthly relative errors: (**a**) Training stations; (**b**) Testing stations.

Supplementary Materials: Supplementary Data are available at https://doi.org/10.17632/9b5tt79s6y.1.

Author Contributions: Conceptualization, J.-I.P.; methodology, J.-I.P.; software, E.G.-P. and D.G.; validation, E.G.-P., D.G. and J.-I.P.; formal analysis, D.G. and J.-I.P.; investigation, E.G.-P., D.G. and J.-I.P.; resources, J.-I.P.; data curation, E.G.-P. and D.G.; writing—review and editing, J.-I.P.; supervision, J.-I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The geographical and climatic characteristics of the weather stations, the experimental data used in this work, and the resulting error calculations are provided as Supplementary Data at https://doi.org/10.17632/9b5tt79s6y.1.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

a_i	Empirical parameter
E'_n	Normalised centred pattern RMSE $= \sqrt{1+\sigma_{sn}^2-2\sigma_{sn}R}$
H	Global solar irradiation on horizontal surface (kWh/m^2)
H_0	Extraterrestrial global solar irradiation on horizontal surface (kWh/m^2)
KGCC	Köppen−Geiger climate classification
L	Distance to the sea (km)
MBE	Mean bias error (kWh/m^2) $= \sum_{i=1}^{n}(s_i-o_i)/n$
NMBE	Normalised mean bias error $= \left(\sum_{i=1}^{n}(s_i-o_i)/n\right)/\overline{o_i}$
NRMSE	Normalised root mean square error $= \left(\sqrt{\sum_{i=1}^{n}(s_i-o_i)^2/n}\right)/\overline{o_i}$
o_i	Observed value
R^2	Coefficient of determination $= \left[\sum_{i=1}^{n}(s_i-\overline{s_i})(o_i-\overline{o_i})/\sqrt{\sum_{i=1}^{n}(s_i-\overline{s_i})^2\sum_{i=1}^{n}(o_i-\overline{o_i})^2}\right]^2$
RMSE	Root mean square error (kWh/m^2) $= \sqrt{\sum_{i=1}^{n}(s_i-o_i)^2/n}$
RMBE	Relative mean bias error $= \sum_{i=1}^{n}((s_i-o_i)/o_i)/n$
RRMSE	Relative root mean square error $= \sqrt{\sum_{i=1}^{n}((s_i-o_i)/o_i)^2/n}$

s_i	Simulated value
T_m	Mean air temperature (K)
T_max	Maximum air temperature (K)
T_min	Minimum air temperature (K)
T_ref	Reference air temperature (K)
z	Elevation above sea level (m)
ΔT	Temperature difference (K) $= T_\mathrm{max} - T_\mathrm{min}$
ϕ	Latitude (°)
λ	Longitude (°)
σ_o	Standard deviation of experimental data $= \sqrt{\left(\sum_{i=1}^{n}(o_i - \overline{o_i})^2\right)/n}$
σ_s	Standard deviation of simulated data $= \sqrt{\left(\sum_{i=1}^{n}(s_i - \overline{s_i})^2\right)/n}$
σ_{sn}	Normalised standard deviation $= \sigma_s/\sigma_o$

References

1. Solangi, K.H.; Islam, M.R.; Saidur, R.; Rahim, N.A.; Fayaz, H. A review on global solar energy policy. *Renew. Sustain. Energy Rev.* **2011**, *15*, 2149–2163. [CrossRef]
2. Izam, N.S.M.N.; Itam, Z.; Sing, W.L.; Syamsir, A. Sustainable Development Perspectives of Solar Energy Technologies with Focus on Solar Photovoltaic—A Review. *Energies* **2022**, *15*, 2790. [CrossRef]
3. Le, H.P.; Sarkodie, S.A. Dynamic linkage between renewable and conventional energy use, environmental quality and economic growth: Evidence from Emerging Market and Developing Economies. *Energy Rep.* **2020**, *6*, 965–973. [CrossRef]
4. Maka, A.O.M.; Alabid, J.M. Solar energy technology and its roles in sustainable development. *Clean. Energy* **2022**, *6*, 476–483. [CrossRef]
5. Perpiña-Castillo, C.; Batista e Silva, F.; Lavalle, C. An assessment of the regional potential for solar power generation in EU-28. *Energy Policy* **2016**, *88*, 86–99. [CrossRef]
6. Amri, F. Intercourse across economic growth, trade and renewable energy consumption in developing and developed countries. *Renew. Sustain. Energy Rev.* **2017**, *69*, 527–534. [CrossRef]
7. Krishnan, N.; Kumar, K.R.; Inda, C.S. How solar radiation forecasting impacts the utilization of solar energy: A critical review. *J. Clean. Prod.* **2023**, *388*, 135860. [CrossRef]
8. Antoñanzas, F.; Sanz, A.; Martínez-de-Pisón, F.J.; Perpiñán, O. Evaluation and improvement of empirical models of global solar irradiation: Case study northern Spain. *Renew. Energy* **2013**, *60*, 604–614. [CrossRef]
9. Ångström, A. Solar and terrestrial radiation. *Q. J. R. Meteorol. Soc.* **1924**, *50*, 121–125.
10. Prescott, J.A. Evaporation from water surface in relation to solar radiation. *Trans. Roy. Soc. Aust.* **1940**, *64*, 114–118.
11. Wan, K.K.W.; Tang, H.L.; Yang, L.; Lam, J.C. An analysis of thermal and solar zone radiation models using an Angstrom–Prescott equation and artificial neural networks. *Energy* **2008**, *33*, 11115–11127. [CrossRef]
12. Chen, J.-L.; He, L.; Yang, H.; Ma, M.; Chen, Q.; Wu, S.-J.; Xiao, Z.-L. Empirical models for estimating monthly global solar radiation: A most comprehensive review and comparative case study in China. *Renew. Sustain. Energy Rev.* **2019**, *108*, 91–111. [CrossRef]
13. Prieto, J.I.; García, D. Global solar radiation models: A critical review from the point of view of homogeneity and case study. *Renew. Sustain. Energy Rev.* **2022**, *155*, 111856. [CrossRef]
14. Prieto, J.I.; García, D.; Santoro, R. Comparative Analysis of Accuracy, Simplicity and Generality of Temperature-Based Global Solar Radiation Models: Application to the Solar Map of Asturias. *Sustainability* **2022**, *14*, 6749. [CrossRef]
15. Paulescu, M.; Stefu, N.; Calinoiu, D.; Paulescu, E.; Pop, N.; Boata, R.; Mares, O. Ångström–Prescott equation: Physical basis, empirical models and sensitivity analysis. *Renew. Sustain. Energy Rev.* **2016**, *62*, 495–496. [CrossRef]
16. Zang, H.; Xu, Q.; Bian, H. Generation of typical solar radiation data for different climates of China. *Energy* **2012**, *38*, 236–248. [CrossRef]
17. Noorollahi, Y.; Mohammadi, M.; Yousefi, H.; Anvari-Moghaddam, A. A Spatial-Based Integration Model for Regional Scale Solar Energy Technical Potential. *Sustainability* **2020**, *12*, 1890. [CrossRef]
18. Thornton, P.E.; Running, S.W. An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. *Agric. For. Meteorol.* **1999**, *93*, 211–228. [CrossRef]
19. Hargreaves, G.H.; Samani, Z.A. Estimating potential evapotranspiration. *J. Irrig. Drain Eng. ASCE* **1982**, *108*, 225–230. [CrossRef]
20. Richardson, C.W. Weather simulation for crop management models. *Trans. ASAE* **1985**, *28*, 1602–1606. [CrossRef]
21. Prieto, J.I.; Martínez-García, J.C.; García, D. Correlation between global solar irradiation and air temperature in Asturias, Spain. *Sol. Energy* **2009**, *83*, 1076–1085. [CrossRef]
22. Prieto, J.I.; Martínez, J.C.; García, D.; Santoro, R.; Rodríguez, A. *Solar Map of Asturias*; Consorcio de Empresas ARFRISOL: Gijón, Spain, 2009. (In Spanish)

23. Prieto, J.I.; García, D. Modified temperature-based global solar radiation models for estimation in regions with scarce experimental data. *Energy Convers. Manag.* **2022**, *268*, 115950. [CrossRef]
24. Attia, S.; Lacombe, T.; Rakotondramiarana, H.T.; Garde, F.; Roshan, G.R. Analysis tool for bioclimatic design strategies in hot humid climates. *Sustain. Cities Soc.* **2019**, *45*, 8–24. [CrossRef]
25. Praene, J.P.; Malet-Damour, B.; Radanielina, M.H.; Fontaine, L.; Rivière, G. GIS-based approach to identify climatic zoning: A hierarchical clustering on principal component analysis. *Build. Environ.* **2019**, *164*, 106330. [CrossRef]
26. Dailidé, R.; Povilanskas, R.; Méndez, J.A.; Simanavičiūtė, G. A new approach to local climate identification in the Baltic Sea's coastal area. *Baltica* **2019**, *32*, 210–218. [CrossRef]
27. Palacios, J. *Dimensional Analysis*; Lee, P., Roth, L., Eds.; Macmillan: London, UK, 1964.
28. Hargreaves, G.H. *Simplified Coefficients for Estimating Monthly Solar Radiation in North America and Europe*; Utah State University: Logan, UT, USA, 1994.
29. Kaushik, S. Clustering | Introduction, Different Methods, and Applications (Updated 2023). Available online: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering (accessed on 29 June 2023).
30. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *WIREs Data Mining Knowl. Discov.* **2011**, *2*, 86–97. [CrossRef]
31. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
32. Rasamoelina, A.D.; Adjailia, F.; Sinčák, P. A Review of Activation Function for Artificial Neural Network. In Proceedings of the IEEE 18th World Symposium on Applied Machine Intelligence and Informatics, Herl'any, Slovakia, 23–25 January 2020.
33. Gueymard, C.A. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. *Renew. Sustain. Energy Rev.* **2014**, *39*, 1024–1034. [CrossRef]
34. Ritter, A.; Muñoz-Carpena, R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *J. Hydrol.* **2013**, *480*, 33–45. [CrossRef]
35. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **2001**, *106*, 7183–7192. [CrossRef]
36. Adaramola, M.S. Estimating global solar radiation using common meteorological data in Akure, Nigeria. *Renew. Energy* **2012**, *47*, 38–44. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.