mathematics

MDPI

Article

Global Solar Irradiation Modelling and Prediction Using Machine Learning Models for Their Potential Use in Renewable Energy Applications

David Puga-Gil [1], Gonzalo Astray [1,*], Enrique Barreiro [2], Juan F. Gálvez [2] and Juan Carlos Mejuto [1]

[1] Universidade de Vigo, Departamento de Química Física, Facultade de Ciencias, 32004 Ourense, España
[2] Universidade de Vigo, Departamento de Informática, Escola Superior de Enxeñaría Informática, 32004 Ourense, España
* Correspondence: gastray@uvigo.es; Tel.: +34-988-387-000

Abstract: Global solar irradiation is an important variable that can be used to determine the suitability of an area to install solar systems; nevertheless, due to the limitations of requiring measurement stations around the entire world, it can be correlated with different meteorological parameters. To confront this issue, different locations in Rias Baixas (Autonomous Community of Galicia, Spain) and combinations of parameters (month and average temperature, among others) were used to develop various machine learning models (random forest -RF-, support vector machine -SVM- and artificial neural network -ANN-). These three approaches were used to model and predict (one month ahead) monthly global solar irradiation using the data from six measurement stations. Afterwards, these models were applied to seven different measurement stations to check if the knowledge acquired could be extrapolated to other locations. In general, the ANN models offered the best results for the development and testing phases of the model, as well as for the phase of knowledge extrapolation to other locations. In this sense, the selected ANNs obtained a mean absolute percentage error (MAPE) value between 3.9 and 13.8% for the model development and an overall MAPE between 4.1 and 12.5% for the other seven locations. ANNs can be a capable tool for modelling and predicting monthly global solar irradiation in areas where data are available and for extrapolating this knowledge to nearby areas.

Keywords: machine learning; random forest; support vector machine; artificial neural network; global solar irradiation; modelling; prediction

MSC: 62P30; 68T05; 62M20; 62M45

Citation: Puga-Gil, D.; Astray, G.; Barreiro, E.; Gálvez, J.F.; Mejuto, J.C. Global Solar Irradiation Modelling and Prediction Using Machine Learning Models for Their Potential Use in Renewable Energy Applications. *Mathematics* **2022**, *10*, 4746. https://doi.org/10.3390/math10244746

Academic Editors: Urmila Diwekar and Debangsu Bhattacharyya

Received: 7 November 2022
Accepted: 6 December 2022
Published: 14 December 2022

1. Introduction

Solar radiation influences all of Earth's processes and presents an essential function in human activity development [1]. This energy source is the principal source for life on Earth and is the most plentiful renewable energy source [2]. As reported by dos Santos et al. (2021) [3], global solar irradiation is an important variable for applications related to the dimensioning of photovoltaic cells, and solar concentrators, among others. This property can be used to approximate and comprehend other linked components such as diffuse solar irradiation, among others [3]. Solar irradiation data are essential to know the performance of solar energy systems [4] and, according to dos Santos et al. (2021) [3], for different applications, it is necessary to assess the solar resource through the data available on ground level. Global solar irradiation is the most relevant variable in renewable energy applications to model the dimension and photovoltaic systems [5].

Global solar irradiation can be carried out by a solarimetric station [3]. Nevertheless, according to Meenal & Selvakumar (2018) [6], it is almost impossible to install solar radiation measuring instruments (to measure solar radiation components) all over the world

due to their price and complicacy of measurements [6]. Due to this, different empirical techniques can be used to correlate solar irradiation and diverse meteorological parameters such as relative humidity, air temperature, etc. [5].

Different factors, such as the high costs of fossil fuels, climate change and global warming, have led to a shift from traditional energy sources to more renewable ones [7]. Additionally, taking into account the European energy crisis unleashed due to the recent war in Ukraine, it is necessary to search for alternatives for energy production to decrease the dependence on fossil sources such as oil or gas. Due to this, it is understood that global solar irradiation determination in the different European territories could be extremely important to plan and carry out different energy generation systems using energy from the sun. In the specific case of Spain, it presents an important dependence on fossil fuels [8]. In this energetic context, it would be interesting to take advantage of the country's favourable climatic conditions for the high production of photovoltaic solar energy (due to its high number of hours of solar radiation in comparison with other European countries) to reduce energy bills [8].

According to the data reported by Fernández-González et al. (2020) [8], Spanish solar photovoltaic production is centred on the territories located in the southern half of the country, which are the areas that received the most radiation, and also present a large amount of surface available for the installation of solar farms. The opposite case within the Spanish territory occurs is the Autonomous Community of Galicia, which is located in the northwest area of the Iberian Peninsula and is one of the climatic zones with the least solar radiation; however, despite this, Galicia is a suitable area for solar installations [9].

Considering all of the above, it seems clear that for the implementation of a solar production system in Galicia, it would be necessary to know the solar irradiation that affects each location in this autonomous community. According to Prieto et al. (2009), to devise of the different solar energy conversion systems, it is required to have reliable monthly average measurements of solar irradiation (it must also be taken into account that meteorological stations with large series of solar irradiation measurements are not widely distributed and that in areas of complex orography, their spatial interpolation is difficult) [10]. To solve data shortages, different techniques have been used to estimate global solar irradiation [3]; in fact, these authors report that these methods can be classified as empirical, regression, machine learning and models that use satellite images. Related to the machine learning (ML) group, machine learning can be defined as a section of artificial intelligence (AI) dedicated to the study of algorithms that can learn autonomously, straight from the input information [11]. There is a broad range of ML models; however, we will focus on the three used in this research, that is, random forest (RF), support vector machine (SVM) and artificial neural network (ANN), which can be used in different fields such as:

- In environmental science to predict metal immobilisation remediation through the use of biochar amendment in soil using different variables such as biochar characteristics or soil physiochemical properties, among others [12], or to predict different important parameters in the Mediterranean Sea [13].
- In medicine to predict the existence of residual cancer after hysterectomy [14] or to help in the diagnosis of COVID-19 using computed tomography images [15].
- In agricultural science in industrial hemp crops to optimise in vitro germination and growth indices [16].

In relation to this research, ML models can be used for different goals. For example, ML models may help to predict monthly global solar irradiation in the interior of the Autonomous Community of Galicia using only three measurement stations and subsequently checking its adjustments at two stations [17]. ML models can also be used to estimate hourly global solar radiation using ANN and SVM models, among others, being the multilayer feed-forward neural network the best developed model [18]. Another interesting study was carried out by Takilalte et al. (2022) [19] to forecast global solar irradiance at different resolutions using a nonlinear autoregressive neural network, RF and support vector regression models in two locations in Algeria. According to the authors, the RF

model shows the best performance [19]. A deep long short-term memory (LSTM) network can be used to predict global solar irradiation one hour ahead using satellite-derived data [20]. On the other hand, de O. Santos et al. (2022) [21] carried out a study to forecast solar irradiance using a heterogeneous ensemble dynamic selection model based on seven methods (multilayer perceptron neural network, support vector regression and RF, among others). The model presents superior overall accuracy (in different error metrics) compared to the single models [21]. Zahraoui et al. (2022) reported different research works, and analysed ML algorithm use in microgrids for short-term temperature and solar irradiation forecasting [22]. Machine learning approaches can be employ to predict very short-term solar irradiation and determine the output power of photovoltaic generators using spatiotemporal variables [23] or to develop models capable of detecting, in real time, the existence of clouds that can potentially block sunrays [24]. Even ANN models can be used to forecast the maintenance in a wind farm [7]. Finally, the possibility of energy consumption forecasting is an interesting fact that could help in the development of energy enterprises' planning strategies or national energy policies [25]. In this sense, Zeng et al. (2017) developed an adaptive differential evolution ANN model that allows reliable prediction of the consumption of this resource.

From another point of view, in a completely independent field from that of photovoltaic energy production, the study area in which this research is carried out is called Rías Baixas, a part of the Galician Rías (which is located on the northern border of the Iberian–Canary Current upwelling system), which has important mussel production [26]. Additionally, Fuentes-Santos et al. (2016) reported that mussel culture has significant commercial value [27,28]. According to the authors, solar irradiance can affect water temperature and food availability, so this variable can have an important function in the prediction of the settlement cycle, starting and ending a month ahead [26].

In the present research work, the modelling and prediction of monthly global solar irradiation are addressed in Rias Baixas (Autonomous Community of Galicia, Spain), an area where solar energy generation has not been exploited strongly. In this sense, the Rías Baixas area also stands out for being an area in which wind farms are not widely located (see maps of Asociación Eólica de Galicia [29] or Rexistro Eólico de Galicia [30]). Due to this fact, it has been thought that the development of tools that allow the modelling and prediction of monthly global solar irradiation could be an interesting way to encourage the placement of panels and solar farms in said area, to fill the gap in the demand for renewable energy.

It also seems clear that the use of prediction models (capable of predicting electrical production or irradiation levels in advance) is necessary for the field of energy production due to its difficult integration into the electrical distribution network. Therefore, this research aimed to model monthly behaviour and to predict monthly global solar irradiation (MGSI) one month ahead, through the application of three different ML models (RF, SVM and ANN).

The models will be developed using data from six measurement stations located in the Rias Baixas, Autonomous Community of Galicia (Spain). Later, with the best selected models, the knowledge acquired will be generalised to seven different stations located in the area. On one hand, the different approaches that will be developed to model the behaviour of monthly global solar irradiation will be an interesting tool capable of determining the solar irradiation conditions in different locations. On the other hand, the approaches developed to predict the monthly global solar irradiation would be useful tools to make a forecast of the conditions one month ahead, which will give an idea to the companies that operate solar farms of possible future electricity production. This work is a brief of the final degree project on environmental sciences developed by Puga-Gil (2022) [31].

2. Materials and Methods

2.1. Study Area

To carry out this study, 13 measurement stations located in the Rías Baixas area of the Galician coast (northwest area of the Iberian Peninsula) were selected (Table 1). These stations were selected taking into account the amount of data available for the development of the different models proposed.

Table 1. Measurement stations used in this research. The six stations on the left correspond to the stations used to train, validate and query the different models. The seven stations on the right correspond to the stations used to study the generalisation of the best selected models.

Stations to Train, Validate and Query	Stations to Generalise Knowledge
Castro Vicaludo (Oia)	Vigo-Campus (Vigo)
Illas Cies (Vigo)	Ponte Caldelas (Ponte Caldelas)
Ons (Bueu)	Cabo Udra (Bueu)
Sálvora-Pazo (Ribeira)	Sanxenxo (Sanxenxo)
Rebordelo (Cotobade)	Porto de Vigo (Vigo)
Fornelos de Montes (Fornelos de Montes)	O Viso (Redondela)
	Lourizán (Pontevedra)

2.2. Database

The database used in this research was obtained from the MeteoGalicia website [32]. The time scale of the downloaded data was monthly, which facilitated the computational work of convergence of the models by greatly reducing the amount of data.

The used input variables were (i,ii) latitude and longitude -WGS84 (EPSG:4326)-, (iii) station altitude in m, (iv) average temperature (°C), (v) average relative humidity (%), (vi) rainfall (L/m^2), (vii) insolation (%) and (viii) hours of sunshine (hours). As the output variable, and the aim of this research, monthly global solar irradiation (MGSI) expressed in 10 $kJ/(m^2 \cdot day)$ was used.

The data belonging to the six measurement stations were divided into three groups. The first group was destined to develop the different models (this group, the training group, is comprised of the data collected between the years 2006 and 2013, both inclusive). The second group of data (validation group, 2014–2018) was used for validation and to select the best model for each kind of model. Finally, the third group (2019–2021) was destined to query each selected model. The data belonging to the other seven stations were used to check the generalisation of the knowledge acquired by the best model previously selected.

The database was debugged to eliminate all the months that lacked any of the variables or presented incorrect data variables (erroneous, suspicious and extrapolated data, etc.).

2.3. Model Implementation

2.3.1. Variable Combinations

In this research, three different types of model were developed: (i) RF, (ii) SVM and (iii) ANN. Different input variable combinations were used to model and predict the MGSI and $MGSI_{+1}$ (monthly global solar irradiation one month ahead), respectively. Block-one is formed by the ML approaches developed to model the monthly global solar irradiation, while Block-two consists of models developed to predict the monthly global solar irradiation one month ahead (Table 2).

Within each block, the models present two different types of input variable combination. Type-1 corresponds to the models developed using input variables: (i–iii) latitude, longitude, altitude, (iv) month, (v) average temperature, (vi) average relative humidity and (vii) rainfall. The use of these variables is justified by the fact that they are frequently measured at most measurement stations; hence, the models that are being developed in this block are those that, in theory, should be easy to implement with the abundance of data. Type-2 corresponds to models for which the variables (average temperature, average rela-

tive humidity and rainfall) were replaced by two variables that have a high correlation (for the training group) with monthly global solar irradiation, that is, insolation (0.753, similar to the correlation of average temperature and MGSI, 0.741) and hours of sunshine (0.936).

Table 2. Variables combination that had been used to develop the different kinds of model used in this research. The output variable is indicated in brown shading and the input variables are indicated in dark brown shading. Adapted from Puga-Gil (2022) [31].

Output Variable		Input Variables								
MGSI	$MGSI_{+1}$	Latitude	Longitude	Altitude	Month	Avg. Temperature	Avg. Relative Humidity	Rainfall	Insolation	Hours of Sunshine
		Block-one, Type-1								
		Block-one, Type-2								
		Block-two, Type-1								
		Block-two, Type-2								

2.3.2. Models Implemented

Random Forest Models

An RF method is an ensemble learning approach which uses a certain number of decision trees as a base learner [33].

According to Kubat (2017) [34], decision trees are based on using a set of rules that compose a tree-like directed graph. The rules are extracted from the features detected in the training set data and, for each new example, those rules are applied to predict to which class it belongs. It starts from the root, and an evaluation of some kind of condition is applied at the step between the nodes of the tree, thus determining the path that the sample follows until it is classified into a class when it reaches a terminal node [34].

However, these decision trees have the problem of overfitting. According to Amro et al. (2021), this is a general problem in machine learning algorithms and can be understood as a classifier presenting good training classification accuracy but being unsuccessful in generalising well with unseen instances [35]. In this sense, random forest is a technique based on decision predictors that significantly reduces the problem of overfitting [36] by creating collections of randomised decision trees, and training each one with different data samples from the training set so that none of them can see the total training samples and each one is trained with different samples of the same problem [37].

RF is commonly used for regression and classification tasks [33,38]. According to different authors [21,39,40], this kind of model was introduced by Breiman [36]. RF is an improvement over bagging classification trees [41].

RF's trees are based on binary partitioning trees where each node is divided into other two nodes, from the root node to the terminal nodes (that are not split) [21].

When RF is used in regression mode, the forecasted value is the average of the predicted values from the trees [40] (Figure 1).

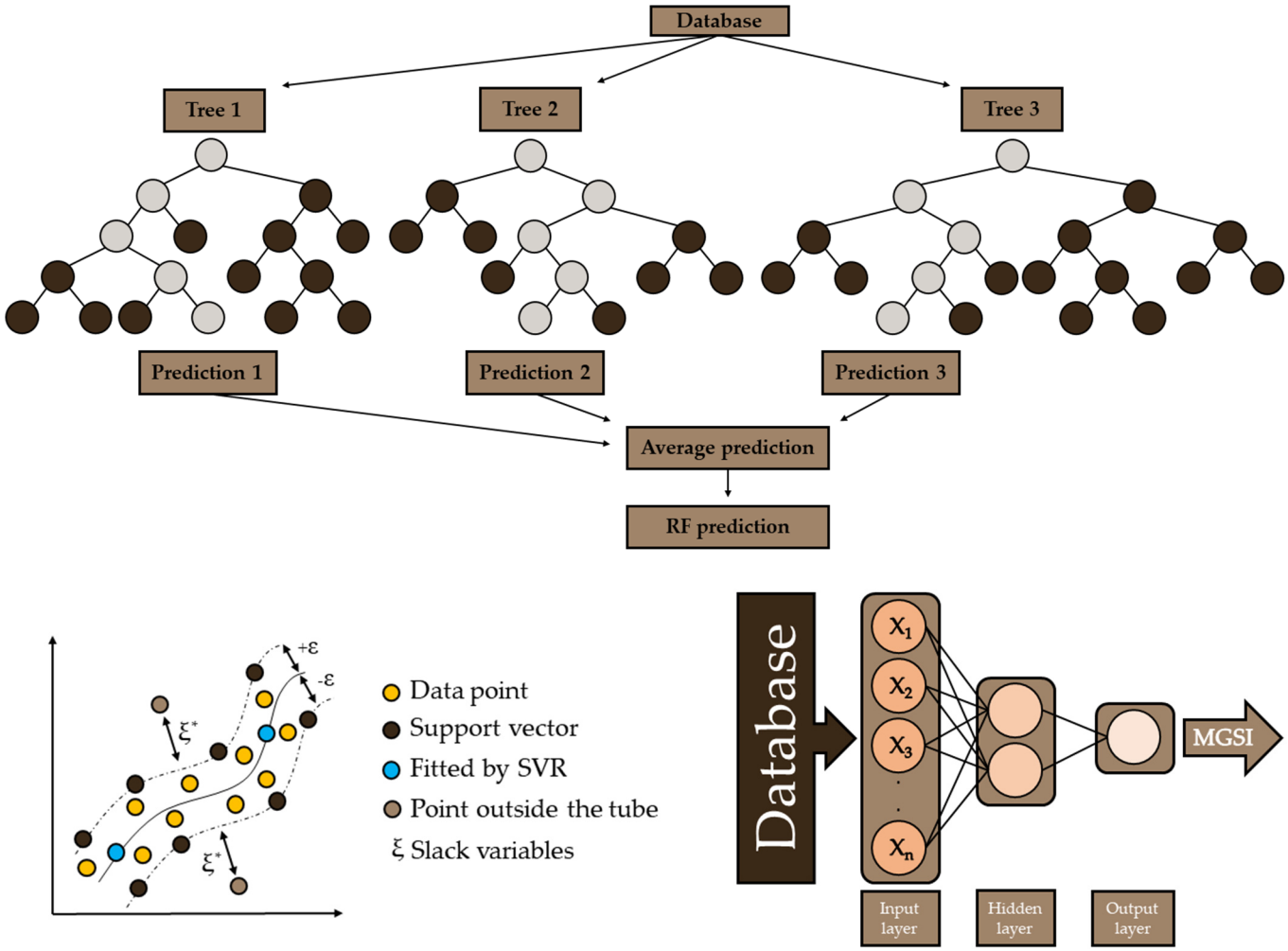

Figure 1. (**top**): scheme of a random forest (RF inspired on Karijadi & Chou (2022) [42]; (**left**): a support vector regression (SVR inspired on Taghizadeh-Mehrjardi et al. (2017) [43]); (**right**): an artificial neural network (ANN inspired on Moldes et al. (2016) [44]).

The RF models applied in this research were implemented using different parameter combinations: the number of trees (between 1 and 100 studied with 99 steps—linear scale), maximum depth (between 1 and 100 with 99 steps—linear scale), prepruning (true or false) and using the least square criterion.

Support Vector Machine Models

According to Tanveer et al. (2022) [45], SVM was introduced in 1995 by Cortes & Vapnik (1995) [46]. This technique is based on finding decision hyperplanes to define the decision boundary to categorise data points into two classes [45]. According to Cervantes et al. (2020), different hyperplanes can correctly separate the input data; nevertheless, there are infinite hyperplanes, so it is necessary to find the separation hyperplane with a maximum margin, or said in another way, the objective of SVM is obtain, using the training set, a surface that maximises the margin to separate various classes [47]. According to Tanveer et al. (2022) [45], SVM can also be used for regression purposes (called support vector regression (SVR) [48]). According to Basak et al. (2007) [48], this technique was proposed by Vapnik et al. (1997) [49] (Figure 1). SVM presents a remarkable generalisation capacity and good discriminative ability, which has allowed this type of model to have caught the attention of different research communities in recent years [47].

This kind of model is characterised by a wide possibility of combining parameters for the development of the models. In this sense, to facilitate the model development processes

and reduce their computational cost, only the combination of SVM types, γ and C were taken into account.

In this case, the combinations used were ε-SVR and ν-SVR for the SVM type, γ was studied between $\approx 2^{-20}$ and 2^{8} (28 steps with a linear and logarithmic scale) and C was studied between $\approx 2^{-10}$ and 2^{20} (30 steps with a linear and logarithmic scale). The RapidMiner operator applied the libsvm learner developed by Chang & Lin (2011, 2021) [50,51]. The values for γ and C were an extension of the ranges proposed by Hsu et al. (2016) in "*A practical guide to support vector classification*" [52]. SVM models were also studied using the normalised and non-normalised database. Normalisation was carried out in the interval [−1,1] for the data of the training phase (firstly for the input variables and later also for both the input and output variables), and this normalisation was applied to all data sets. Once the best model was selected, the data were de-normalised to compare all the models with each other.

Artificial Neural Network Models

An artificial neural network is a computational model designed to mimic the biological nervous system [53]. Another possible definition would be that an ANN is a nonlinear statistical technique that can be used to model the intricate relationship between data (called input and output data) [12]. ANN can be considered a "universal approximator," that is, ANN can be used to resolve problems related to pattern recognition, categorisation, data generation and approximation [54]. Artificial neural network architecture (also called topology) is based on the biological neural network's structure and function [55]. Within ANNs, the multilayer perceptron (MLP) is probably the most commonly used ANN [43].

The ANN topology consists of different layers (input, hidden and output (Figure 1) that are made up of neurons [56,57]. These layers are assigned different functions: the input layer receives the data from the database and sends them to the hidden layer, the output layer provides a prediction and the hidden layer, or layers, carries out, together with the output layer, the data processing [43]. Between the neurons of the different layers, there is a connection (synapse) that corresponds with a weight value that depends on the connection strength [56]. The learning process consists of finding the correlation between the input data and the output feedback through variation in the associated weights [58]. In our specific case, the input layer receives the MeteoGalicia meteorological data and during the training, the information flows through the artificial neural network, modifying the internal weights and biases to obtain the lowest possible error.

The neural networks developed in this research were implemented with a single hidden layer. The number of units in the hidden layer was studied between 1 and $2n + 1$ (n corresponds with the number of neurons in the first neural network layer), while the training cycles were studied between 1 and 131,072 (17 steps in linear and logarithmic scales) and decay (true or false).

2.4. Statistics

To evaluate the different models implemented, the following statistical parameters were used: (i) the root mean square error (RMSE), (ii) the mean absolute percentage error (MAPE) and (iii) the correlation coefficient (r). The best model within each model approach was chosen using the RMSE in the validation phase. Then, the best models were tested using the reserved querying cases.

2.5. Equipment and Software Used

The models developed in this research were implemented using an Intel ® Core ™ i9-10900 2.80 GHz with 64 GB of RAM and an Intel ® Core ™ i7-8700 3.2 GHz with 32 GB of RAM.

Data were assembled using Microsoft Excel 2016. The RF, SVM and ANN models were developed using an educational and a free version of RapidMiner Studio 9.10.001 soft-

ware (RapidMiner GmbH, Dortmund, Germany). Figures were drawn using the software SigmaPlot v. 13.0 (Palo Alto, CA, USA).

3. Results and Discussion

3.1. Approaches to Modelling Monthly Global Solar Irradiation (Block-One)

Table 3 shows the best models developed to model (Block-one), and predict (Block-two), the monthly global solar irradiation according to each variable combination (Type-1 and Type-2).

Table 3. Adjustments obtained for each selected ML approach developed to model the MGSI (Block-one) and to predict the MGSI one month ahead (Block-two) according to each input variable's configuration (Type). The best models (in terms of RMSE in the validation phase) within each Block and Type are indicated in bold. RMSE corresponds with the root mean square error, r corresponds with the correlation coefficient and subscripts T, V and Q correspond to the phases; training, validation and querying, respectively. Adapted and data from Puga-Gil (2022) [31].

Model	$RMSE_T$	r_T	$RMSE_V$	r_V	$RMSE_Q$	r_Q
			Block-one, Type-1			
RF	75.8	0.994	163.0	0.977	202.9	0.974
SVM	144.9	0.977	154.2	0.982	203.4	0.979
ANN	**155.6**	**0.980**	**121.4**	**0.985**	**154.0**	**0.982**
			Block-one, Type-2			
RF	50.5	0.997	94.7	0.992	184.7	0.969
SVM	87.9	0.991	73.3	0.995	166.6	0.977
ANN	**96.3**	**0.990**	**68.7**	**0.995**	**109.3**	**0.992**
			Block-two, Type-1			
RF	128.0	0.983	215.1	0.954	250.2	0.943
SVM	202.4	0.956	219.3	0.950	249.2	0.939
ANN	**192.2**	**0.960**	**198.3**	**0.959**	**228.7**	**0.953**
			Block-two, Type-2			
RF	125.8	0.983	225.1	0.947	262.8	0.933
SVM	179.6	0.964	207.7	0.956	293.8	0.914
ANN	**189.9**	**0.961**	**195.9**	**0.960**	**215.2**	**0.957**

Block-one Type-1 corresponds to the models developed that used the following input variables to model the MGSI: (i–iii) latitude, longitude, altitude, (iv) month, (v) average temperature, (vi) average relative humidity and (vii) rainfall. In this sense, as can be seen, given the results obtained from the RMSE in the validation phase, the best model is the ANN model. The RF model and the SVM model show higher values of RMSE (163.0 10 kJ/(m^2·day) and 154.2 10 kJ/(m^2·day), respectively) compared to the ANN model (121.4 10 kJ/(m^2·day)). Despite the difference in the RMSE values, all the models present a high correlation coefficient for the validation phase (between 0.977 and 0.985) with acceptable MAPE values (between 7.2% and 10.3%). Given these adjustments, it can be said that the three models chosen to determine the monthly global solar irradiation present good behaviour for the validation phase. This good behaviour is also observed in the training phase where the models show RMSE values between 75.8 10 kJ/(m^2·day) and 155.6 10 kJ/(m^2·day). The good behaviour of these models was tested using the querying data. As can be seen in Table 3, the ANN model offers the best adjustments with an RMSE value of 154.0 10 kJ/(m^2·day), while the other two models present higher RMSE values, 202.9 10 kJ/(m^2·day) and 203.4 10 kJ/(m^2·day), for the RF and the SVM model, respectively.

Figure 2 shows the real and modelled MGSI values obtained by the different selected ANN models.

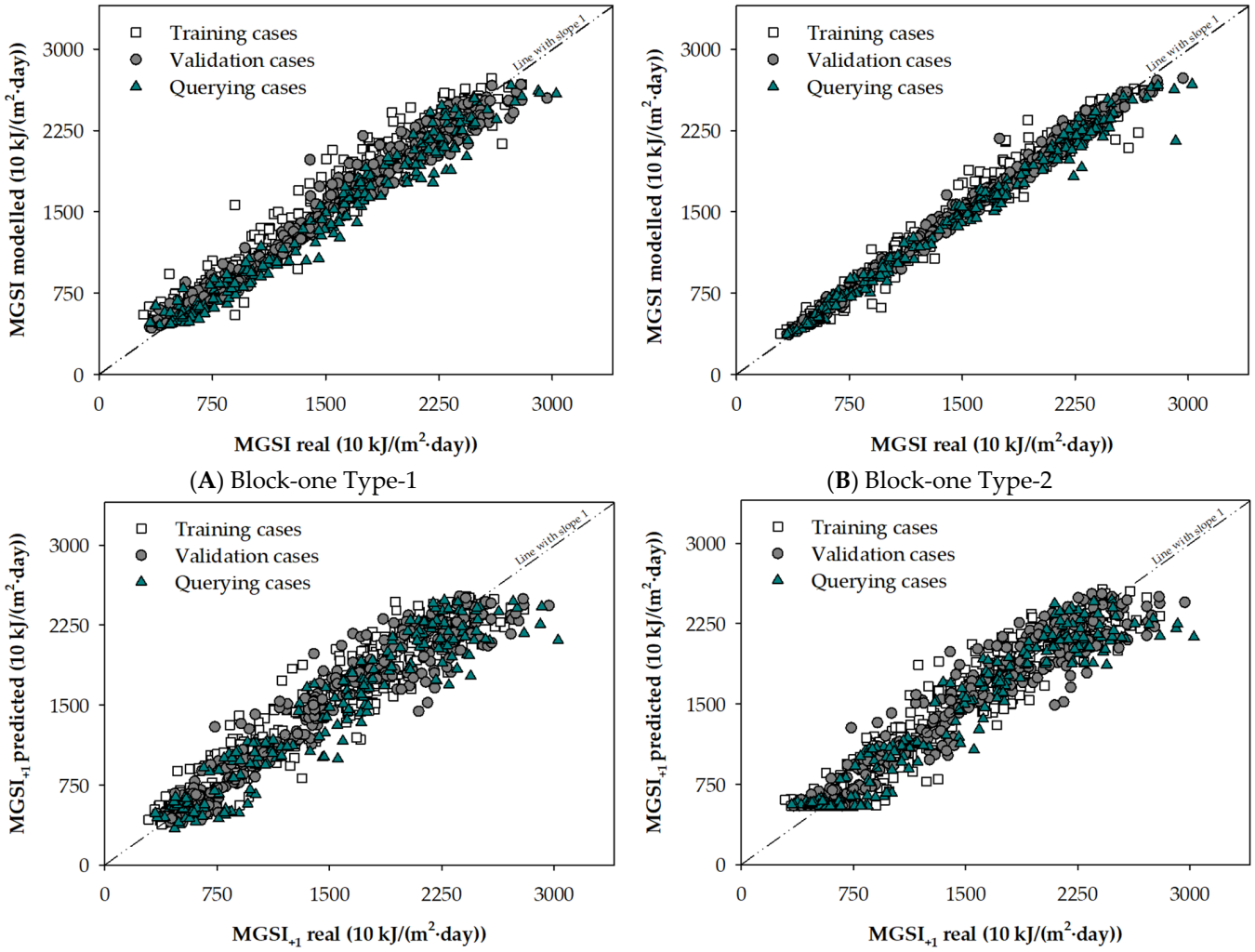

Figure 2. Scatter plot for the real and modelled/predicted values of MGSI for the training (white squares), the validation (grey dots) and the querying cases (cyan triangles) for each ANN selected model using the combination of the variables to model the MGSI (Block-one Type-1 (**A**) and Block-one Type-2 (**B**)), and to predict MGSI one month ahead (Block-two Type-1 (**C**) and Block-two Type-2 (**D**)). Dashed and dotted line is the line of slope one. Adapted from Puga-Gil (2022) [31].

Figure 2A shows that in the training phase (white dots) some cases are far from the line of slope one (dashed and dotted line). For this phase, the model ANN presents an MAPE value of 11%. This percentage is based on the fact that there are several points that present a high percentage error value; an example of this is a case (left side of the graph) whose real value is 898 10 kJ/(m^2·day) but which is predicted by the model to have a value of 1564.9 10 kJ/(m^2·day), that is, an overestimation of 74.3%. The most important error, in terms of percentage error, occurs for a case for which the true value is 463 10 kJ/(m^2·day) but which is predicted by the model to have a value of 928.1 10 kJ/(m^2·day), that is, an overestimation of 100.5%. In the validation phase, this behaviour changes and the MAPE drops to 7.2%. Figure 2A shows that the cases are closer to the line of slope one. Even so, some cases are slightly further away from the line as with one case (near the centre of the graph) whose real value is 1395 10 kJ/(m^2·day) and which the model predicts to have a value of 1980.1 10 kJ/(m^2·day), that is, an overestimation of 41.9%. For the querying phase, the cases are a little further from the line of slope one but remain at an acceptable MAPE value of 9.3%. In this case, the largest errors in terms of percentage error are centred in the lower area of the graph, such as a case for which the real value is 371 10 kJ/(m^2·day) and

which is predicted by the model to have a value of 633.2 10 kJ/(m^2·day), which represents an overestimation of 70.7%.

Finally, considering all the adjustments obtained by the ANN model and the low MAPE for the querying phase (9.3%), it can be stated that the selected artificial neural network model is an appropriate tool to model the MGSI.

Block-one Type-2 corresponds to the models developed using six input variables ((i–iii) latitude, longitude, altitude, (iv) month, (v) insolation and (vi) hours of sunshine) to model the MGSI. As mentioned above, the substitution of average temperature, average relative humidity and rainfall by the insolation and hours of sunshine variables should allow substantial improvement in the adjustments of the best selected models. In this case, as can be seen, the root mean square error for the validation phase shows that the ANN model provides the best results (Table 3). The differences are tighter than in the case of models belonging Type-1, at least when the ANN model is compared to the SVM model concerning root mean square error (68.7 10 kJ/(m^2·day) and 73.3 10 kJ/(m^2·day), respectively). The random forest model shows the highest value of RMSE (94.7 10 kJ/(m^2·day)). This random forest model is also the model with the worst adjustments in the querying phase. Once again, in this phase (querying phase) the best model selected based on the root mean square error in the validation phase, the ANN model, shows the lowest RMSE (109.3 10 kJ/(m^2·day)) and MAPE (4.9%) values and the highest correlation coefficient (0.992).

Figure 2B shows the real and modelled MGSI values obtained by the selected ANN model for Block-one Type-2. The dispersion of the cases around the line of slope one is much lower than for the previously chosen model. This is probably due to the high correlation that exists between the MGSI and the hours of sunshine. This high correlation allows the model to obtain better predictions, and therefore, the dispersion decreases. In the training phase, some cases in the upper part of the graph are far from the line of slope one; these correspond to three cases with underestimated values between 14.0 and 19.6%. An extreme case occurs for a value of 959 10 kJ/(m^2·day), which presents an underestimation of 35.7% (616.4 10 kJ/(m^2·day)). For the validation phase, a decrease in the dispersion is observed, with only one case moving away from the line of slope one. This case shows a real value of 1747 10 kJ/(m^2·day) and a prediction of 2175.9 10 kJ/(m^2·day), which represents an overestimation of 24.6%. Finally, in the querying phase, all cases are closer to the line of slope one, although three cases can be seen that differ slightly, highlighting a case whose real value (2917 10 kJ/(m^2·day)) is underestimated by the model by 26.2% (2151.9 10 kJ/(m^2·day)).

To concluded, it can be said that the ANN model developed using insolation and hours of sunshine (models Type-2) is a model that offers good results in all the analysed phases and it is a suitable tool to model the MGSI. Nevertheless, these models are more complicated to apply because these two variables are not widely measured at all stations.

Following the above, it can be verified, when comparing models Type-1 and Type-2, that the latter improve the model's adjustments for monthly global solar irradiation. The good performance of these models can be seen in the RMSE and MAPE values. The improvements are evident for all phases; for example, in the validation phase, it can be seen that for the best ANN models, the RMSE value falls from 121.4 10 kJ/(m^2·day) to 68.7 10 kJ/(m^2·day), which corresponds to a decrease of 43.4% in the value of the RMSE. The same behaviour can be seen in the querying phase where the error falls from 154.0 10 kJ/(m^2·day) to 109.3 10 kJ/(m^2·day) (a decrease of 29.0%). Given the results, it can be said that any of the two models of artificial neural networks selected are valid tools for use in the prediction of monthly global solar irradiation.

3.2. Approaches to Predicting the Monthly Global Solar Irradiation One Month Ahead (Block-Two)

Table 3 presents the best Type-1 models developed to predict the monthly global solar irradiation one month ahead according to each variable combination and model kind. As can be seen, the model that provides the best behaviour for the validation phase (based on the root mean square error) is the ANN model, showing an RMSE value of

198.3 10 kJ/(m^2·day), which corresponds to an MAPE of 11.3%. On the other hand, the RF and the SVM model show worse results than the ANN model. In fact, and based on the root mean square error, both models show RMSE values of (215.1 10 kJ/(m^2·day) and 219.3 10 kJ/(m^2·day), respectively). Similarly to the models of Block-one Type-1 for the validation phase, the correlation coefficient also remains high (0.950–0.959) although it drops slightly. This behaviour is also observed for the training phase, where the models show RMSE values between 128.0 10 kJ/(m^2·day) (RF model) and 202.4 10 kJ/(m^2·day) (SVM model). As expected, for the querying phase, the adjustments go down, showing values of RMSE between 228.7 10 kJ/(m^2·day) (ANN model) and 250.2 10 kJ/(m^2·day) (RF model).

As can be seen in Table 3, all the Block-two Type-1 models (to predict the MGSI one month ahead) offer worse adjustments than the Block-one Type-1 models. This worsening in the adjustments is observable for all the phases of the models, varying, regarding RMSE, from 22.5% for the SVM model in the querying phase to 68.7% for the RF model in the training phase. For the best developed model, the ANN model, the deteriorations in the adjustments (training (23.5%), validation (63.3%) and querying (48.5%)) are also important. The worsening of the results is logical because it is more difficult to predict behaviour in advance (in this case, a month ahead) than to model the behaviour in the same month.

Figure 2C shows the real and predicted MGSI values obtained by the selected ANN model for Block-two Type-1. As expected, the chosen artificial neural network model has a much higher dispersion than its counterpart, Block-one Type-1, chosen to model the MGSI. For all the phases of model development, the dispersion is greater than for the Block-one Type-1 model. As previously stated, it is evident that predicting the variable of interest a month ahead is more complicated than modelling its behaviour in the same month; hence, there is greater dispersion. Due to this, all the development phases of the model present a higher MAPE value, which varies between 11.3% for the validation phase and 13.8% for the querying phase. For the training phase, there is a case which presents an overestimation of 82.7% (484 10 kJ/(m^2·day) vs. 884.4 10 kJ/(m^2·day)). The validation phase presents an MAPE value lower than those of the training and querying phase; in fact, for this phase, the maximum percentage of error for a given case is 76.3% (overestimation) which corresponds to a real value of 735 10 kJ/(m^2·day) that is predicted as 1295.7 10 kJ/(m^2·day). Finally, for the querying phase, where the model has a higher MAPE value (13.8%), a case with a higher percentage of error is predicted to have a value of 487.0 10 kJ/(m^2·day) when its real value is 897 10 kJ/(m^2·day), that is, the model underestimated this case by 45.7%.

Type-2 corresponds to the models that use a smaller number of variables, that is, (i–iii) latitude, longitude, altitude, (iv) month, (v) insolation and (vi) hours of sunshine. As can be seen in Table 3, and considering the RMSE for the validation phase, the model with the worst results for the validation phase is the RF model that shows an RMSE value of 225.1 10 kJ/(m^2·day), which corresponds to an MAPE of 14.1%. This model presents a higher RMSE for the querying phase, 262.8 10 kJ/(m^2·day), which corresponds to 15.1%. The SVM model obtains an RMSE of 207.7 10 kJ/(m^2·day) for the validation phase, which is slightly lower than that presented by the RF model. However, this model, despite presenting better results for the validation phase than the RF model, presents, for the querying phase, the highest RMSE value (293.8 10 kJ/(m^2·day)) of all the developed models. Finally, the best model corresponds, once again (based on the RMSE value in the validation phase), to the artificial neural network model. This ANN model shows the lowest RMSE (195.9 10 kJ/(m^2·day)) and MAPE (11.6%) values and the highest correlation coefficient (0.960).

The Block-two Type-2 models (to predict the MGSI one month ahead) offer worse adjustments than the Block-one Type-2 models (to model monthly global solar irradiation). This worsening in the adjustments is observable for all the phases of the ANN model, varying, in terms of RMSE, between 42.3% for the RF model in the querying phase and 185.2% for the ANN models in the validation phase.

Figure 2D shows the real and predicted MGSI values obtained by the selected ANN model for Block-two Type-2. Once again, as expected, the ANN model presents higher dispersion than its counterpart (Block-one Type-2) chosen to model the MGSI. All the development phases of the model present a higher MAPE value, which varies between 11.6% for the validation phase and 13.3% for the training phase. For the training phase, the largest prediction error occurs for a case whose real value is 293 10 kJ/(m^2·day) and is overestimated by the model by 106.0% (603.7 10 kJ/(m^2·day)). In the validation phase, the point with the highest dispersion is the one whose real value is 735 10 kJ/(m^2·day) and it is predicted by the model to have a value of 1278.0 10 kJ/(m^2·day), which corresponds to an overestimation of 73.9%. Finally, for the querying phase, three points present a percentage error greater than 60% (overestimation), specifically with values ranging from 66.6% to 60.8%. These three cases correspond to the three lowest values of the variable for the querying phase.

It can be concluded that considering the MAPE value for the validation and querying phases, the ANN model could be suitable to predict the MGSI value one month ahead.

Previously, it had been seen for Block-one that the models that were developed using the insolation and the hours of sunshine (Type-2) presented better results than the models that did not use them (Type-1). For Block-two, it can also be seen that the models developed for Type-2 present better results, except for the Type-1 RF model in the validation and querying phase and the SVM model in the querying phase. In Block-two Type-2, the improvements in the SVM and ANN models are smaller than the improvements observed in the models corresponding to Block-one. The improvement in the validation phase for the ANN model is only 1.2%; however, in the querying phase, the ANN model improves the RMSE value by around 5.9%. Given these results, it can be concluded that in the models to predict MGSI one month ahead, the use of insolation and hours of sunshine does not cause significant improvement in the predictions of the ANN model.

3.3. Generalisation at Different Locations

After the selection of the best ML models in the previous section, it is time to verify that the models developed for the six stations can be successfully applied to other locations using the seven reserved stations which have not been used for training, validation, and querying. The adjustments for the best selected approaches applied to these new seven stations are listed in Table 4.

The best model for Block-one Type-1 is the ANN model, which shows an RMSE value for the $Q_{2–8}$ stations of 151.3 10 kJ/(m^2·day), while the other two models present values of 165.1 10 kJ/(m^2·day) (RF model) and 166.3 10 kJ/(m^2·day) (SVM model), with MAPE values of 10.3% and 11.1 and 11.9%, respectively. The ANN model shows the best adjustments (in terms of RMSE) for five of the seven stations analysed. This model is overcome by the SVM model for station Q_2 (Vigo-Campus (Vigo)) and by the RF and SVM models in Q_8 (Lourizán (Pontevedra)), respectively (Table 4 and Figure 3A). On one hand, the SVM model shows a lower RMSE value for station Q_2 (141.1 10 kJ/(m^2·day)) than the RF (203.4 10 kJ/(m^2·day)) and ANN (143.6 10 kJ/(m^2·day)) models. On the other hand, the RF model shows a lower RMSE value for station Q_8 (144.4 10 kJ/(m^2·day)) than the SVM (154.7 10 kJ/(m^2·day)) and ANN (169.3 10 kJ/(m^2·day)) models.

Given the results obtained for each station, and in all the stations in general, it can be said that the ANN to model the MGSI using the input variables (i–iii) latitude, longitude, altitude, (iv) month, (v) average temperature, (vi) average relative humidity and (vii) rainfall (Block-one Type-1) is a good modelling tool that provides an overall mean absolute percentage error around 10.3%, which is considered an acceptable error.

The best model for Block-one Type-2 is the ANN model, which shows an RMSE value for the $Q_{2–8}$ stations of 73.0 10 kJ/(m^2·day), which corresponds with an MAPE value of 4.1%, while the RF model presents a root mean square error of 110.7 10 kJ/(m^2·day) and the SVM models present an RMSE of 102.1 10 kJ/(m^2·day). In this case, the ANN model shows the best adjustments for six of the seven stations analysed, with the Q_2 station

(Vigo-Campus (Vigo)) being the only station where the adjustments obtained by the ANN model are improved (Table 4 and Figure 3B). In this case, the SVM model presents better adjustments, with an RMSE of 66.7 10 kJ/(m^2·day), which corresponds to an MAPE of 4.6%.

Table 4. Adjustments obtained for each selected model developed to model the MGSI (Block-one) and to predict the MGSI one month ahead (Block-two) according to each input variable's configuration (Type). RMSE corresponds to the root mean square error and r corresponds with the correlation coefficient for each external querying station: Q_2—Vigo-Campus (Vigo), Q_3—Ponte Caldelas (Ponte Caldelas), Q_4—Cabo Udra (Bueu), Q_5—Sanxenxo (Sanxenxo), Q_6—Porto de Vigo (Vigo), Q_7—O Viso (Redondela) and Q_8—Lourizán (Pontevedra). The best results (in terms of RMSE) for each station are indicated in bold. Data from Puga-Gil (2022) [31].

	Q_2		Q_3		Q_4		Q_5		Q_6		Q_7		Q_8	
Model	**RMSE**	**r**	**RMSE**	**r**	**RMSE**	**r**	**RMSE**	**r**	**RMSE**	**r**	**RMSE**	**r**	**RMSE**	**r**
Block-one, Type-1														
RF	203.4	0.973	157.5	0.979	155.5	0.984	163.2	0.977	182.8	0.976	145.6	0.980	**144.4**	**0.978**
SVM	**141.1**	**0.982**	149.6	0.979	143.1	0.986	168.7	0.981	192.4	0.974	196.9	0.980	154.7	0.980
ANN	143.6	0.982	**127.0**	**0.984**	**124.4**	**0.989**	**154.3**	**0.986**	**167.6**	**0.979**	**143.7**	**0.986**	169.3	0.976
Block-one, Type-2														
RF	110.6	0.991	97.9	0.992	165.0	0.981	105.5	0.994	128.9	0.986	109.0	0.988	89.4	0.993
SVM	**66.7**	**0.996**	79.5	0.995	158.3	0.976	80.8	0.996	133.4	0.991	119.1	0.992	97.7	0.995
ANN	87.6	0.996	**48.0**	**0.998**	**73.2**	**0.997**	**66.3**	**0.997**	**102.9**	**0.992**	**72.9**	**0.994**	**57.1**	**0.997**
Block-two, Type-1														
RF	261.0	0.948	242.7	0.940	206.5	0.966	191.5	0.963	221.8	0.962	199.9	0.960	**180.3**	**0.964**
SVM	268.8	0.944	246.4	0.934	**175.4**	**0.970**	193.2	0.961	247.8	0.964	201.5	0.956	191.3	0.963
ANN	**241.1**	**0.957**	**224.4**	**0.948**	179.2	0.970	**177.2**	**0.967**	**191.7**	**0.970**	**185.8**	**0.962**	184.8	0.965
Block-two, Type-2														
RF	**236.9**	**0.947**	259.4	0.931	221.0	0.956	198.2	0.960	**203.5**	**0.967**	**196.8**	**0.963**	186.9	0.961
SVM	381.0	0.948	286.3	0.942	262.5	0.967	259.5	0.964	300.7	0.963	317.3	0.952	283.3	0.962
ANN	245.0	0.959	**220.8**	**0.954**	**204.7**	**0.963**	**190.9**	**0.966**	230.2	0.959	203.0	0.957	**176.2**	**0.965**

With all these, the ANN to model the MGSI using the input variables (i–iii) latitude, longitude, altitude, (iv) month, (v) insolation and (vi) hours of sunshine is a good modelling approach that provides an overall MAPE of 4.1%.

The best model for Block-two Type-1 is, once again, the ANN model, which presents an RMSE value for all the stations of 198.1 10 kJ/(m^2·day), which corresponds with an MAPE value of 12.0%. The other two models present worse results: 212.8 10 kJ/(m^2·day) for the RF model and 218.3 10 kJ/(m^2·day) for the SVM model. This increase in the RMSE values is also reflected in the MAPE values, which, in this case, are 13.8% and 15.2%, respectively. The ANN model shows the best adjustments for five of the seven stations, with the only two stations where this does not occur being Q_4 (Cabo Udra (Bueu) and Q_8 (Lourizán (Pontevedra)) (Table 4 and Figure 3C). In Q_4, the SVM model presents a better RMSE (175.4 10 kJ/(m^2·day)) than the RMSE obtained by the ANN model (179.2 10 kJ/(m^2·day)), while in Q_8, the RF model presents an RMSE value of 180.3 10 kJ/(m^2·day) against the 184.8 10 kJ/(m^2·day) presented by the ANN model.

The ANN model intended to predict the MGSI one month ahead that used the input variables (i–iii) latitude, longitude, altitude, (iv) month, (v) average temperature, (vi) average relative humidity and (vii) rainfall is a good prediction model that provides results with acceptable MAPE values between 10.6 and 13.5%.

The best model for Block-two Type-2 is the ANN model, which presents an RMSE value for all the stations of 207.8 10 kJ/(m^2·day), which corresponds with an MAPE value of 12.5%. However, in this case, the random forest model predicts three stations with better adjustments than the ANN model (Table 4 and Figure 3D). This occurs for the Q_2 (Vigo-Campus (Vigo)), Q_6 (Porto de Vigo (Vigo)) and Q_7 (O Viso (Redondela)) stations, where

the random forest model presents RMSE values of 236.9, 203.5 and 196.8 10 kJ/(m^2·day), against 245.0, 230.2 and 203.0 10 kJ/(m^2·day) obtained by the ANN model.

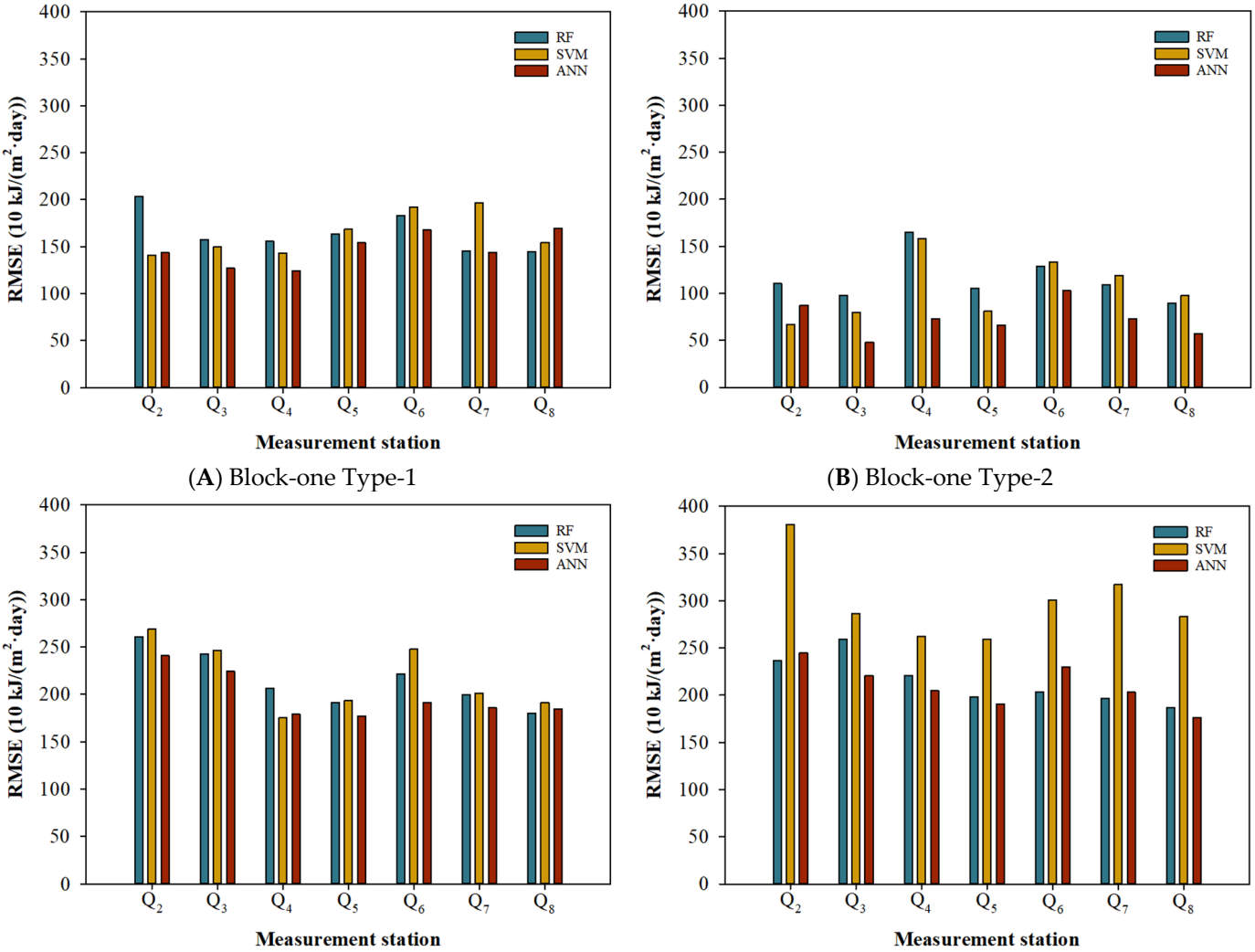

Figure 3. Root mean square error (RMSE) for each selected model based on the measurement station (Q_2—Vigo-Campus (Vigo), Q_3—Ponte Caldelas (Ponte Caldelas), Q_4—Cabo Udra (Bueu), Q_5—Sanxenxo (Sanxenxo), Q_6—Porto de Vigo (Vigo), Q_7—O Viso (Redondela), and Q_8—Lourizán (Pontevedra)) and the combination of the variables (Block-one Type-1 (**A**) and Block-one Type-2 (**B**)) and to predict MGSI one month ahead (Block-two Type-1 (**C**) and Block-two Type-2 (**D**)). Data from Puga-Gil (2022) [31].

This ANN model to predict the MGSI one month ahead using the input variables (i–iii) latitude, longitude, altitude, (iv) month, (v) insolation and (vi) hours of sunshine, is a good prediction model that provides good results with acceptable MAPE values between 11.1 and 13.8%.

Figures 4 and 5 show the different time series for the seven stations studied in this phase (Q_2–Q_8).

These series show the different cycles that exist throughout the time series, where the maximum and minimum MGSI correspond to the summer and winter seasons, respectively. This can also be seen, for example, in the Sanxenxo (Sanxenxo) station, as one of these cycles seems to overlap; this is due to the debug task mentioned in Section 2.2. The real

values of MGSI are shaded in cyan while the modelling/predictions made by the different artificial neural networks are shown as a black line.

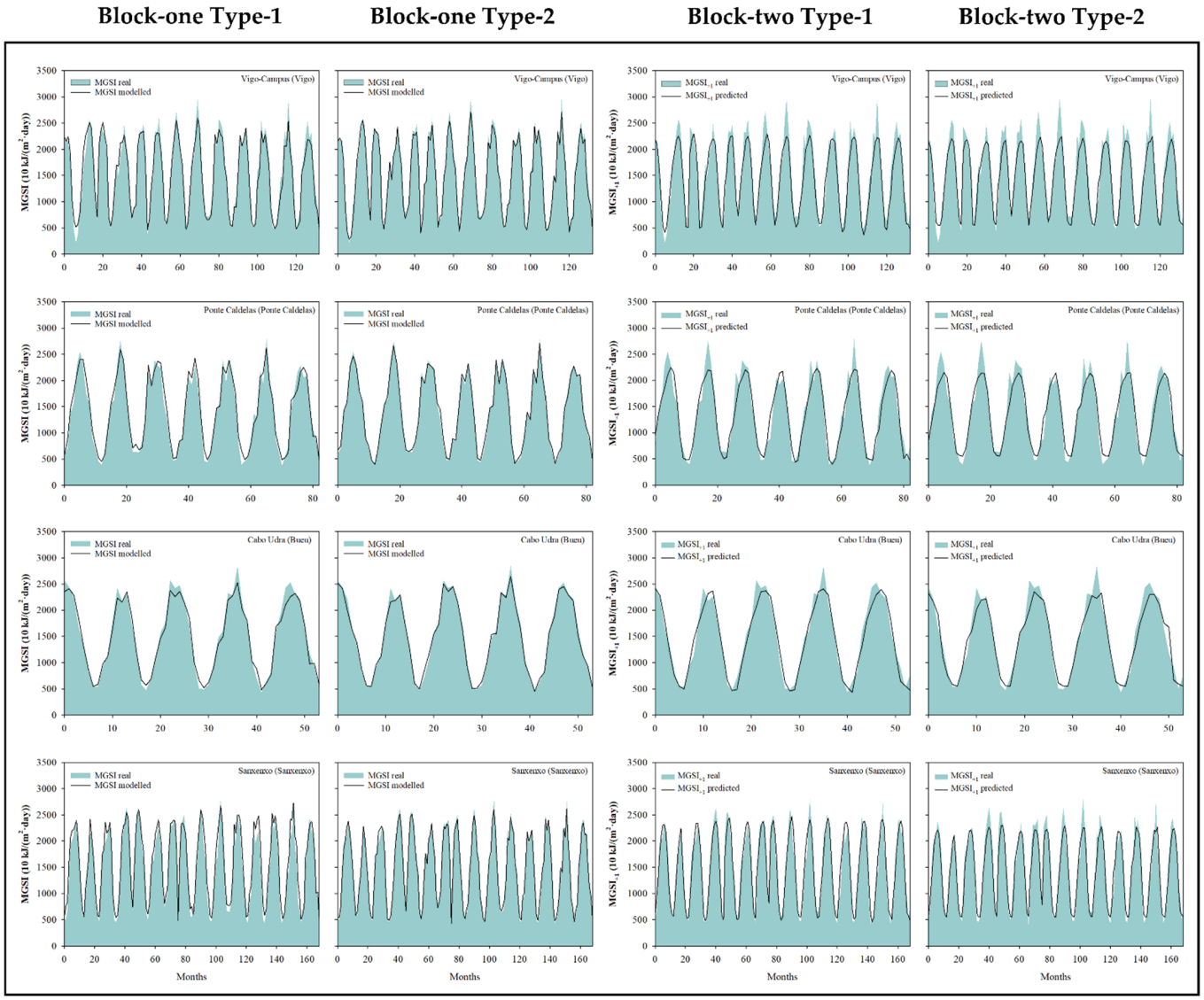

Figure 4. Real and modelled/predicted MGSI values for each measurement station according to the combination of the used variables to model the MGSI (Block-one Type-1 and Block-one Type-2) and to predict MGSI one month ahead (Block-two Type-1 and Block-two Type-2) for the stations: Q_2—Vigo-Campus (Vigo), Q_3—Ponte Caldelas (Ponte Caldelas), Q_4—Cabo Udra (Bueu) and Q_5—Sanxenxo (Sanxenxo). The cyan shade is the real value of MGSI and the black line is the modelled/predicted MGSI value by each selected ANN model. Data from Puga-Gil (2022) [31].

Looking at the ANNs to model the behaviour of the MGSI, the two columns of figures on the left (Block-one Type-1 and Type-2) show how the modelling of both ANNs captured the behaviour of the GMSI for each station. This is seen in the overlap of the black line in practically the entire length of the time series, except for some points where the prediction slightly deviates from the shaded area. This fact can be seen in some MGSI peaks for Block-one Type-1 at different stations such as the Q_2 (Vigo-Campus (Vigo), Q_4 (Cabo Udra (Bueu)) and Q_6 (Porto de Vigo (Vigo)) stations. The ANN developed to model the behaviour of the variable using the Type-2 combination (Block-one Type-2) presents fewer mismatches in the modelling line. This fact is not only appreciated by looking at the charts shown in Figures 4 and 5 but it has also been observed in the statistical data

where the Type-2 model substantially improved the Type-1 model (73.0 10 kJ/(m^2·day) vs. 151.3 10 kJ/(m^2·day), respectively).

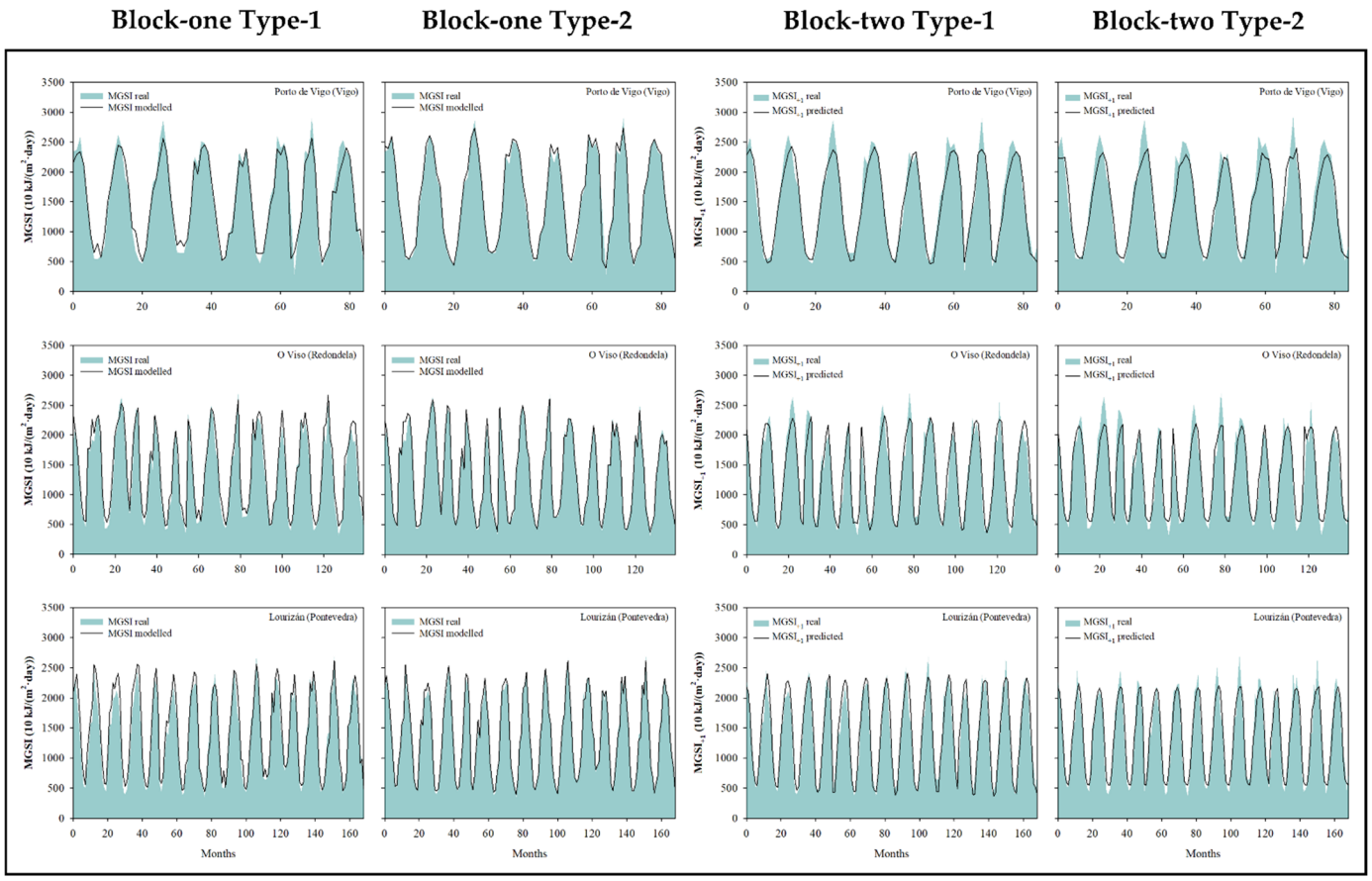

Figure 5. Real and modelled/predicted MGSI values for each measurement station according to the combination of the used variables to model the MGSI (Block-one Type-1 and Block-one Type-2) and to predict MGSI one month ahead (Block-two Type-1 and Block-two Type-2) for the stations: Q_6—Porto de Vigo (Vigo), Q_7—O Viso (Redondela) and Q_8—Lourizán (Pontevedra)). The cyan shade is the real value of MGSI and the black line is the modelled/predicted MGSI value by each selected ANN model. Data from Puga-Gil (2022) [31].

On the other hand, it can be seen that in the ANN models developed to predict the MGSI one month ahead (Block-two Type-1 and Type-2, the two columns of figures on the right) the black line does not overlap the real series as much as the ANNs designed to model the MGSI (Figures 4 and 5). This fact is appreciable for both kinds of variable selection (Type-1 and Type-2); furthermore, between these two models, there is no appreciable difference. Considering the statistical data, this assumption can be confirmed, because the Type-1 model presents an overall RMSE for $Q_{2–8}$ of 198.1 10 kJ/(m^2·day) (MAPE of 12.0%), while the Type-2 ANN model presents an RMSE of 207.8 10 kJ/(m^2·day) (MAPE of 12.5%); the difference between them is not significant.

It seems clear that the selected artificial neural network approaches can perform good modelling and predictions of monthly global solar irradiation using easily obtainable variables at the different measurement stations. This fact constitutes an advantage over other models reported in the literature that include other input variables that need previous work and adaptation prior to their use, such as extraterrestrial radiation measurements, satellite images, etc.

When the results obtained in this research are compared with similar studies carried out in other locations, it can be said that the models developed here present adequate

adjustments. Wang et al. (2016) carried out different neural network techniques to predict solar radiation using different variables such as air temperature, relative humidity, etc. The different neural models obtained RMSE errors between 1.94 MJ/(m^2·day) and 3.29 MJ/(m^2·day) for the testing phase [59]. Similar research was developed by Diez et al. (2020) to predict the horizontal daily global solar irradiation in a region close to the study area of this article, in which they developed models of artificial neural networks using a series of eight-year data. The authors used different input combination variables (horizontal daily global solar irradiation one/two days before, number of the day of the year, and others) to determine the horizontal daily global solar irradiation one day after [1]. Diez et al. (2020) concluded that the RMSE value obtained by their best and worst neural network model varied between 3.770 MJ/(m^2·d)) and 4.261 MJ/(m^2·d)). Both studies differ slightly from ours (by the input variables used, the output variable to be determined or the study time window), but despite this, it can give us frame of reference for the results obtained in this research. The best ANN model developed in this research to model monthly global solar irradiation obtained RMSE values of 73.0 10 kJ/(m^2·day) and 151.3 10 kJ/(m^2·day) for the overall $Q_{2–8}$ phase. On the other hand, the RMSE values to predict the monthly global solar irradiation one month ahead were 198.1 10 kJ/(m^2·day) and 207.8 10 kJ/(m^2·day) in the overall $Q_{2–8}$ phase.

Another interesting study reported in the literature is the one developed by Huang et al. (2021), who modelled daily and monthly solar radiation using different machine learning algorithms. The authors selected different variables (average relative humidity, minimum pressure and average soil temperature, among others) to develop their models [60]. The RMSE for monthly values were between 1.131 MJ/m^2 and 1.580 MJ/m^2 for the total period studied. In this case, the models developed in this research present similar behaviour (151.3 10 kJ/(m^2·day)) in the upper zone of RMSE, although the lower zone presents better results (73.0 10 kJ/(m^2·day)).

On the other hand, to the authors' knowledge, the use of the models outside the studied areas could lead to a loss of predictive power. This should be verified by obtaining similar parameters and their application in different areas; this was not the aim of the study in this research, but would be interesting to investigate in future research. This clarification will always remain clear, as happens in the work presented by Walch et al. (2019), where they develop a machine learning model to determine annual solar irradiation on rooftops. In the discussion and conclusion sections of this interesting research, the authors report that the ML model is only applicable to sites with comparable latitudes to Switzerland [61].

In conclusion, the selected ANNs models can be used in Rias Baixas (Autonomous Community of Galicia, Spain) to model and predict monthly global solar irradiation. The predictions obtained by these models can be used to choose the location of solar systems and solar farms, or even serve as input data for the prediction of solar energy production. Finally, these models must always be applied to areas with similar characteristics to those that have been used for their development.

3.4. Final Remarks—Analysis of the Models via a Database Update

Once all the models had been developed, and their generalisation power in the other locations had been checked, MeteoGalicia made a change in the location coordinates of one of the stations, specifically, the Illas Cíes station. The Google Maps tool [62] allows us to verify that both locations are located very close to each other, at around 310 m. Due to this small discrepancy, we decided to study how this change could affect the predictions reported by the best selected models.

The new adjustments showed that the three different phases, where the Illas Cíes station is included (T, V and Q), present a minimum variation, in terms of RMSE, between −0.168% and 0.063%. Given these results, it can be concluded that the chosen developed models are practically unaffected in their adjustments by the database updates.

4. Conclusions

Three kinds of ML model (RF, SVM and ANN) were developed in this research to model and predict MGSI. The best MLP models generally work with appropriate MAPE values. Given the results provided in Table 3, it can be said that, in general, the models that provide the best adjustments for all phases are the artificial neural network models, both for the models of Block-one and Block-two, as well as for those of Type-1 and Type-2.

The selected neural network models have shown their ability to model and predict monthly global solar irradiation, not only in the stations used to develop the models (MAPE values between 3.9 and 13.8%—Table 3) but also in the overall Q_{2-8} phase (MAPE values between 4.1 and 12.5%). This means that artificial neural network models can be applied to model and predict monthly global solar irradiation in locations in which data are available for the development of models, and that the knowledge acquired from these data can be extrapolated to nearby locations.

As expected, based on the results presented, it is also possible to verify that the developed ML methods to model the MGSI present better results than the ML ones to predict the MGSI one month ahead. Finally, based on the results (Table 4, Figure 3), it can be said that the SVM model presents similar RMSE values to the RF model, except in the case of forecasting the MGSI one month ahead using the Type-2 variable distribution.

ANN models could be an interesting tool to determine which are the best locations to install systems/solar farms; moreover, once they are installed, the models would allow the prediction of the value of monthly global solar irradiation one month ahead, which would enable producers to predict the energy that they could supply.

A notable fact is the evidence that has been found regarding the updating of the station's location coordinates. This modification presents minimal variation in the predictions of the developed models, in terms of RMSE. This allows us to affirm that the developed neural network models could be fault-tolerant.

The models selected in this research could be utilize with good results in the areas under study, that is, in the studied part of Rias Baixas (Autonomous Community of Galicia, Spain). It is understood that these models can be extrapolated to surrounding areas, but always with a loss of modelling/predictive power. Extrapolation to more remote areas would probably lead to unreliable results due to the different relationships established between the input variables used in the model and the variable to be modelled/predicted.

Finally, it is necessary to clarify that all these models could be improved with the use of more measurement stations, new/different input variables, different datasets for training, validation and querying, analyses of different parameter ranges for each model, etc.

Author Contributions: Conceptualisation, G.A. and J.C.M.; methodology, D.P.-G. and G.A.; validation, D.P.-G. and G.A.; formal analysis, D.P.-G. and G.A.; investigation, D.P.-G. and G.A.; writing—original draft preparation, D.P.-G. and G.A.; writing—review and editing, D.P.-G., G.A., E.B., J.F.G. and J.C.M.; visualisation, D.P.-G. and G.A.; supervision, G.A. and J.C.M. 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 used data can be obtained on the website of MeteoGalicia (Consellería de Medio Ambiente, Territorio e Vivenda of Xunta de Galicia) [32].

Acknowledgments: The authors thank MeteoGalicia and the Consellería de Medio Ambiente, Territorio e Vivenda of Xunta de Galicia for providing access to the database used in this research. The authors also thank RapidMiner Inc. for the free and the educational licenses of RapidMiner Studio 9.10.001 software. This work is a summary of the final degree project developed by D.P.-G.

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

Nomenclature

AI	Artificial intelligence
ANN	Artificial neural network
Block-one	Group of ML approaches developed to model the MGSI
Block-two	Group of ML approaches developed to predict the $MGSI_{+1}$
LSTM	Long short-term memory
MAPE	Mean absolute percentage error
MGSI	Monthly global solar irradiation
$MGSI_{+1}$	Monthly global solar irradiation one month ahead
ML	Machine learning
MLP	Multilayer perceptron
Q	Querying phase
Q_2	Querying phase for Vigo-Campus station (Vigo)
Q_3	Querying phase for Ponte Caldelas station (Ponte Caldelas)
Q_4	Querying phase for Cabo Udra station (Bueu)
Q_5	Querying phase for Sanxenxo station (Sanxenxo)
Q_6	Querying phase for Porto de Vigo station (Vigo)
Q_7	Querying phase for O Viso station (Redondela)
Q_8	Querying phase for Lourizán station (Pontevedra)
r	Correlation coefficient
RF	Random forest
RMSE	Root mean square error
SVM	Support vector machine
T	Training phase
Type-1	Variable combination Type-1
Type-2	Variable combination Type-2
V	Validation phase

References

1. Diez, F.J.; Navas-Gracia, L.M.; Chico-Santamarta, L.; Correa-Guimaraes, A.; Martínez-Rodríguez, A. Prediction of Horizontal Daily Global Solar Irradiation Using Artificial Neural Networks (ANNs) in the Castile and León Region, Spain. *Agronomy* **2020**, *10*, 96. [CrossRef]
2. Kambezidis, H.D. The Solar Radiation Climate of Greece. *Climate* **2021**, *9*, 183. [CrossRef]
3. dos Santos, C.M.; Teramoto, É.T.; de Souza, A.; Aristone, F.; Ihaddadene, R. Several Models to Estimate Daily Global Solar Irradiation: Adjustment and Evaluation. *Arab. J. Geosci.* **2021**, *14*, 286. [CrossRef]
4. Mubiru, J.; Banda, E.J.K.B. Monthly Average Daily Global Solar Irradiation Maps for Uganda: A Location in the Equatorial Region. *Renew. Energy* **2012**, *41*, 412–415. [CrossRef]
5. Yacef, R.; Benghanem, M.; Mellit, A. Prediction of Daily Global Solar Irradiation Data Using Bayesian Neural Network: A Comparative Study. *Renew. Energy* **2012**, *48*, 146–154. [CrossRef]
6. Meenal, R.; Selvakumar, A.I. Assessment of SVM, Empirical and ANN Based Solar Radiation Prediction Models with Most Influencing Input Parameters. *Renew. Energy* **2018**, *121*, 324–343. [CrossRef]
7. Sa'ad, A.; Nyoungue, A.C.; Hajej, Z. An Integrated Maintenance and Power Generation Forecast by ANN Approach Based on Availability Maximization of a Wind Farm. *Energy Reports* **2022**, *8*, 282–301. [CrossRef]
8. Fernández-González, R.; Suárez-García, A.; Álvarez Feijoo, M.Á.; Arce, E.; Díez-Mediavilla, M. Spanish Photovoltaic Solar Energy: Institutional Change, Financial Effects, and the Business Sector. *Sustainability* **2020**, *12*, 1892. [CrossRef]
9. Vázquez Vázquez, M. *Atlas de Radiación Solar de Galicia*; Vázquez Vázquez, M., Ed.; Universidade de Vigo: Vigo, Spain, 2005; ISBN 84-609-7101-5.
10. 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]
11. Bertolini, M.; Mezzogori, D.; Neroni, M.; Zammori, F. Machine Learning for Industrial Applications: A Comprehensive Literature Review. *Expert Syst. Appl.* **2021**, *175*, 114820. [CrossRef]
12. Sun, Y.; Zhang, Y.; Lu, L.; Wu, Y.; Zhang, Y.; Kamran, M.A.; Chen, B. The Application of Machine Learning Methods for Prediction of Metal Immobilization Remediation by Biochar Amendment in Soil. *Sci. Total Environ.* **2022**, *829*, 154668. [CrossRef]
13. Astray, G.; Soto, B.; Barreiro, E.; Gálvez, J.F.; Mejuto, J.C. Machine Learning Applied to the Oxygen-18 Isotopic Composition, Salinity and Temperature/Potential Temperature in the Mediterranean Sea. *Mathematics* **2021**, *9*, 2523. [CrossRef]
14. Ganguli, R.; Franklin, J.; Yu, X.; Lin, A.; Heffernan, D.S. Machine Learning Methods to Predict Presence of Residual Cancer Following Hysterectomy. *Sci. Rep.* **2022**, *12*, 2738. [CrossRef] [PubMed]

15. Al-Areqi, F.; Konyar, M.Z. Effectiveness Evaluation of Different Feature Extraction Methods for Classification of COVID-19 from Computed Tomography Images: A High Accuracy Classification Study. *Biomed. Signal Process. Control* **2022**, *76*, 103662. [CrossRef]
16. Aasim, M.; Katırcı, R.; Akgur, O.; Yildirim, B.; Mustafa, Z.; Nadeem, M.A.; Baloch, F.S.; Karakoy, T.; Yılmaz, G. Machine Learning (ML) Algorithms and Artificial Neural Network for Optimizing in Vitro Germination and Growth Indices of Industrial Hemp (*Cannabis sativa* L.). *Ind. Crops Prod.* **2022**, *181*, 114801. [CrossRef]
17. Martinez-Castillo, C.; Astray, G.; Mejuto, J.C. Modelling and Prediction of Monthly Global Irradiation Using Different Prediction Models. *Energies* **2021**, *14*, 2332. [CrossRef]
18. Guher, A.B.; Tasdemir, S.; Yaniktepe, B. Effective Estimation of Hourly Global Solar Radiation Using Machine Learning Algorithms. *Int. J. Photoenergy* **2020**, *2020*, 8843620. [CrossRef]
19. Takilalte, A.; Harrouni, S.; Mora, J. Forecasting Global Solar Irradiance for Various Resolutions Using Time Series Models—Case Study: Algeria. *Energy Sources Part A Recover. Util. Environ. Eff.* **2022**, *44*, 1–20. [CrossRef]
20. Benamrou, B.; Ouardouz, M.; Allaouzi, I.; Ben Ahmed, M. A Proposed Model to Forecast Hourly Global Solar Irradiation Based on Satellite Derived Data, Deep Learning and Machine Learning Approaches. *J. Ecol. Eng.* **2020**, *21*, 26–38. [CrossRef]
21. de O. Santos, D.S.; de Mattos Neto, P.S.G.; de Oliveira, J.F.L.; Siqueira, H.V.; Barchi, T.M.; Lima, A.R.; Madeiro, F.; Dantas, D.A.P.; Converti, A.; Pereira, A.C.; et al. Solar Irradiance Forecasting Using Dynamic Ensemble Selection. *Appl. Sci.* **2022**, *12*, 3510. [CrossRef]
22. Zahraoui, Y.; Alhamrouni, I.; Mekhilef, S.; Basir Khan, M.R. Machine Learning Algorithms Used for Short-Term PV Solar Irradiation and Temperature Forecasting at Microgrid. In *Applications of AI and IOT in Renewable Energy*; Shaw, R.N., Ghosh, A., Mekhilef, S., Balas, V.E., Eds.; Academic Press: Cambridge, MA, USA, 2022; pp. 1–17. ISBN 978-0-323-91699-8.
23. Rodríguez, F.; Martín, F.; Fontán, L.; Galarza, A. Ensemble of Machine Learning and Spatiotemporal Parameters to Forecast Very Short-Term Solar Irradiation to Compute Photovoltaic Generators' Output Power. *Energy* **2021**, *229*, 120647. [CrossRef]
24. Nespoli, A.; Niccolai, A.; Ogliari, E.; Perego, G.; Collino, E.; Ronzio, D. Machine Learning Techniques for Solar Irradiation Nowcasting: Cloud Type Classification Forecast through Satellite Data and Imagery. *Appl. Energy* **2022**, *305*, 117834. [CrossRef]
25. Zeng, Y.-R.; Zeng, Y.; Choi, B.; Wang, L. Multifactor-Influenced Energy Consumption Forecasting Using Enhanced Back-Propagation Neural Network. *Energy* **2017**, *127*, 381–396. [CrossRef]
26. Fuentes-Santos, I.; Labarta, U.; Álvarez-Salgado, X.A.; Fernández-Reiriz, M.J. Solar Irradiance Dictates Settlement Timing and Intensity of Marine Mussels. *Sci. Rep.* **2016**, *6*, 29405. [CrossRef] [PubMed]
27. Díaz, C.; Figueroa, Y.; Sobenes, C. Seasonal Effects of the Seeding on the Growth of Chilean Mussel (*Mytilus edulis platensis*, d'Orbigny 1846) Cultivated in Central Chile. *Aquaculture* **2014**, *428–429*, 215–222. [CrossRef]
28. Labarta, U.; Fernández-Reiriz, M.J.; Pérez-Camacho, A.; Pérez-Corbacho, E. *Bateeiros, Mar, Mejillón. Una Perspectiva Bioeconómica*; Fundación Caixa Galicia y Centro de Investigación Económica y Financiera (CIEF): Vigo, Spain, 2004; ISBN 84-95491-69-9.
29. Asociación Eólica de Galicia, Mapa de Parques Eólicos En Explotación En Galicia (a Fecha 31/12/2019). Available online: https://www.ega-asociacioneolicagalicia.es/el-sector-en-cifras/ (accessed on 18 November 2022).
30. Xunta de Galicia Consellería de Industria, Rexistro Eólico Galicia v2 PRD. Available online: https://www.arcgis.com/apps/webappviewer/index.html?id=4bae3fad95b6439bacef9d1a316765e9 (accessed on 18 November 2022).
31. Puga-Gil, D. Modelado y Predicción de La Irradiación Solar Global Mensual En La Zona Inferior de Las Rías Baixas Usando Modelos de Aprendizaje Automático. Final Degree Project, Universidade de Vigo, Ourense, Spain, 2022.
32. MeteoGalicia; Consellería de Medio Ambiente Territorio e Vivenda; Xunta de Galicia MeteoGalicia. Available online: https://www.meteogalicia.gal/ (accessed on 22 January 2022).
33. Zhang, L.; Lu, S.; Ding, Y.; Duan, D.; Wang, Y.; Wang, P.; Yang, L.; Fan, H.; Cheng, Y. Probability Prediction of Short-Term User-Level Load Based on Random Forest and Kernel Density Estimation. *Energy Rep.* **2022**, *8*, 1130–1138. [CrossRef]
34. Kubat, M. Decision Trees. In *An Introduction to Machine Learning*; Kubat, M., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 113–135. ISBN 978-3-319-63913-0.
35. Amro, A.; Al-Akhras, M.; El Hindi, K.; Habib, M.; Shawar, B.A. Instance Reduction for Avoiding Overfitting in Decision Trees. *J. Intell. Syst.* **2021**, *30*, 438–459. [CrossRef]
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Biau, G.; Scornet, E. A Random Forest Guided Tour. *TEST* **2016**, *25*, 197–227. [CrossRef]
38. Djandja, O.S.; Salami, A.A.; Wang, Z.-C.; Duo, J.; Yin, L.-X.; Duan, P.-G. Random Forest-Based Modeling for Insights on Phosphorus Content in Hydrochar Produced from Hydrothermal Carbonization of Sewage Sludge. *Energy* **2022**, *245*, 123295. [CrossRef]
39. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning: Methods and Applications*; Zhang, C., Ma, Y., Eds.; Springer: Boston, MA, USA, 2012; pp. 157–175. ISBN 978-1-4419-9326-7.
40. Tyralis, H.; Papacharalampous, G.; Langousis, A. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water* **2019**, *11*, 910. [CrossRef]
41. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
42. Karijadi, I.; Chou, S.-Y. A Hybrid RF-LSTM Based on CEEMDAN for Improving the Accuracy of Building Energy Consumption Prediction. *Energy Build.* **2022**, *259*, 111908. [CrossRef]

43. Taghizadeh-Mehrjardi, R.; Neupane, R.; Sood, K.; Kumar, S. Artificial Bee Colony Feature Selection Algorithm Combined with Machine Learning Algorithms to Predict Vertical and Lateral Distribution of Soil Organic Matter in South Dakota, USA. *Carbon Manag.* **2017**, *8*, 277–291. [CrossRef]
44. Moldes, Ó.A.; Morales, J.; Cid, A.; Astray, G.; Montoya, I.A.; Mejuto, J.C. Electrical Percolation of AOT-Based Microemulsions with n-Alcohols. *J. Mol. Liq.* **2016**, *215*, 18–23. [CrossRef]
45. Tanveer, M.; Rajani, T.; Rastogi, R.; Shao, Y.H.; Ganaie, M.A. Comprehensive Review on Twin Support Vector Machines. *Ann. Oper. Res.* **2022**, 1–46. [CrossRef]
46. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
47. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
48. Basak, D.; Pal, S.; Patranabis, D.C. Support Vector Regression. *Neural Inf. Process.–Lett. Rev.* **2007**, *11*, 203–224.
49. Vapnik, V.; Golowich, S.E.; Smola, A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In *Advances in neural information Processing System 9, Proceedings of the 1996 Conference*; Mozer, C., Jordan, M.I., Petsche, T., Eds.; MIT Press: Cambridge, MA, USA, 1997; pp. 281–287.
50. Chang, C.-C.; Lin, C.-J. LIBSVM—A Library for Support Vector Machines. Available online: https://www.csie.ntu.edu.tw/~{}cjlin/libsvm/ (accessed on 17 October 2022).
51. Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
52. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification. Available online: https://www.csie.ntu.edu.tw/~{}cjlin/papers/guide/guide.pdf (accessed on 17 October 2022).
53. Kareem, S.S.; Pathak, Y. Chapter 20—Clinical Applications of Artificial Neural Networks in Pharmacokinetic Modeling. In *Artificial Neural Network for Drug Design, Delivery and Disposition*; Puri, M., Pathak, Y., Sutariya, V.K., Tipparaju, S., Moreno, W., Eds.; Academic Press: Walthan, MA, USA, 2016; pp. 393–405. ISBN 978-0-12-801559-9.
54. Han, K.; Wang, Y. A Review of Artificial Neural Network Techniques for Environmental Issues Prediction. *J. Therm. Anal. Calorim.* **2021**, *145*, 2191–2207. [CrossRef]
55. Sairamya, N.J.; Susmitha, L.; Thomas George, S.; Subathra, M.S.P. Hybrid Approach for Classification of Electroencephalographic Signals Using Time–Frequency Images with Wavelets and Texture Features. In *Intelligent Data Analysis for Biomedical Applications Challenges and Solutions*; Hemanth, D.J., Gupta, D., Emilia Balas, V., Eds.; Academic Press: Cambridge, MA, USA, 2019; pp. 253–273. ISBN 978-0-12-815553-0.
56. Zarra, T.; Galang, M.G.; Ballesteros, F.; Belgiorno, V.; Naddeo, V. Environmental Odour Management by Artificial Neural Network—A Review. *Environ. Int.* **2019**, *133*, 105189. [CrossRef] [PubMed]
57. Abdolrasol, M.G.M.; Hussain, S.M.S.; Ustun, T.S.; Sarker, M.R.; Hannan, M.A.; Mohamed, R.; Ali, J.A.; Mekhilef, S.; Milad, A. Artificial Neural Networks Based Optimization Techniques: A Review. *Electronics* **2021**, *10*, 2689. [CrossRef]
58. Isabona, J.; Imoize, A.L.; Ojo, S.; Karunwi, O.; Kim, Y.; Lee, C.-C.; Li, C.-T. Development of a Multilayer Perceptron Neural Network for Optimal Predictive Modeling in Urban Microcellular Radio Environments. *Appl. Sci.* **2022**, *12*, 5713. [CrossRef]
59. Wang, L.; Kisi, O.; Zounemat-Kermani, M.; Salazar, G.A.; Zhu, Z.; Gong, W. Solar Radiation Prediction Using Different Techniques: Model Evaluation and Comparison. *Renew. Sustain. Energy Rev.* **2016**, *61*, 384–397. [CrossRef]
60. Huang, L.; Kang, J.; Wan, M.; Fang, L.; Zhang, C.; Zeng, Z. Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events. *Front. Earth Sci.* **2021**, *9*, 596860. [CrossRef]
61. Walch, A.; Castello, R.; Mohajeri, N.; Scartezzini, J.-L. A Fast Machine Learning Model for Large-Scale Estimation of Annual Solar Irradiation on Rooftops. In Proceedings of the ISES Solar World Congress 2019 and IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry, Santiago, Chile, 3–7 November 2019; pp. 2201–2210.
62. Google LLC. Google Maps. Available online: https://www.google.es/maps/?hl=es (accessed on 12 July 2022).