Energy Conversion and Management 62 (2012) 1–13

Contents lists available at SciVerse ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

# Comparison of Gene Expression Programming with neuro-fuzzy and neural network computing techniques in estimating daily incoming solar radiation in the Basque Country (Northern Spain)

Gorka Landeras [a], José Javier López [b], Ozgur Kisi [c], Jalal Shiri [d,*]

[a] NEIKER, AB, Basque Country Research Institute for Agricultural Development, Alava, Basque Country, Spain
[b] Department of Projects and Rural Engineering of the Public University of Navarre, Campus de Arrosadía, 31006 Pamplona, Spain
[c] Civil Engineering Department, Faculty of Architecture and Engineering, CanikBasari University, Samsun, Turkey
[d] Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran

ARTICLE INFO

*Article history:*
Received 16 January 2012
Received in revised form 16 March 2012
Accepted 29 March 2012
Available online 25 June 2012

*Keywords:*
Global solar radiation
Artificial intelligence
Gene Expression Programming
Temperature

ABSTRACT

Surface incoming solar radiation is a key variable for many agricultural, meteorological and solar energy conversion related applications. In absence of the required meteorological sensors for the detection of global solar radiation it is necessary to estimate this variable. Temperature based modeling procedures are reported in this study for estimating daily incoming solar radiation by using Gene Expression Programming (GEP) for the first time, and other artificial intelligence models such as Artificial Neural Networks (ANNs), and Adaptive Neuro-Fuzzy Inference System (ANFIS). A comparison was also made among these techniques and traditional temperature based global solar radiation estimation equations. Root mean square error (RMSE), mean absolute error (MAE) RMSE-based skill score ($SS_{RMSE}$), MAE-based skill score ($SS_{MAE}$) and $r^2$ criterion of Nash and Sutcliffe criteria were used to assess the models' performances. An ANN (a four-input multilayer perceptron with 10 neurons in the hidden layer) presented the best performance among the studied models (2.93 MJ $m^{-2}$ $d^{-1}$ of RMSE). The ability of GEP approach to model global solar radiation based on daily atmospheric variables was found to be satisfactory.



## 1. Introduction

Accurate estimations of global solar radiation data are needed for design and development of energy efficient buildings and solar energy conversion (photovoltaic or solar thermal) systems. The correct observation/estimation of surface incoming solar radiation ($R_S$) is also very important for many agricultural and meteorological applications such as evapotranspiration estimation, crop simulation and biomass accumulation models. While most weather stations are provided with sensors for air temperature detection, the presence of sensors necessary for the detection of global solar radiation is not so habitual and the data quality provided by them is sometimes poor [1]. In the last few years there has been an increase in the number of automatic weather stations with electronic sensors although the problem is that they require qualified staff for the maintenance of their sensors, which increases the cost of their use. The realization of studies related with many agricultural and meteorological applications requires long time series of global solar radiation. Therefore the availability of long time series of global solar radiation values it is not very usual.

These problems can be overcome by using different approaches to estimate $R_S$. One of these approaches is based on the modeling of physical relation between maximum and minimum temperatures, and $R_S$. The models based on this approach use daily air temperature range to estimate atmospheric transmissivity. On the one hand, cloud cover decreases the maximum air temperature because of the smaller input of short wave radiation. On the other hand, cloud cover increases the minimum air temperature during night time because of the greater emissivity of clouds compared to clear sky [2]. This approach is often selected by many authors, for its simplicity and the small amount of variables needed. It has been included in many recent studies [3–6].

Another alternative for the estimation of global solar radiation is the utilization of artificial intelligence techniques. Recently, under artificial intelligences (AI) category, the Artificial Neural Networks (ANNs), Adaptive Neuro-Fuzzy Inference System (ANFIS) and Genetic Programming (GP) approaches have been successfully used in a wide range of scientific applications including water resources engineering, agro-hydrology and agro-meteorology [7]. The complete review of all of the published papers of AI applications is beyond the scope of this paper and only studies that are related to global solar radiation simulation are mentioned here.

Since Elizondo et al. [8] formulated a neural network for estimating daily global solar radiation, many authors have used the

* Corresponding author.
*E-mail addresses:* glanderas@neiker.net (G. Landeras), j_shiri2005@yahoo.com (J. Shiri).

0196-8904/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.enconman.2012.03.025

Artificial Neural Networks (ANNs) based global solar radiation estimation approach [9–13]. The application of Adaptive Neuro-Fuzzy Inference System (ANFIS) for estimating global solar radiation has not been very common [14].

The studies mentioned above show the potential of AI techniques as design tools for estimation of global solar radiation. Nevertheless, the survey of the literature by the authors of this paper showed that no study has been carried out to utilize Genetic Programming techniques (such as Gene Expression Programming) for global solar radiation estimation. This provides an impetus for the current work. The aim of the present work was to apply for the first time Gene Expression Programming techniques for the estimation of global solar radiation daily values from routinely observed meteorological data, and to compare their performance with other AI based approaches (Artificial Neural Networks, Adaptive Neuro-Fuzzy Inference Systems) and with empirical equations.

## 2. Models overview

### 2.1. Empirical equations

The empirical models used for this comparative study are the Hargreaves and Samani model (HS), the Mahmood and Hubbard model (MH1) and the debiased Mahmood and Hubbard model (MH2). The HS model was chosen as reference model for its long term proven efficacy and its parsimonious formulation.

Hargreaves and Samani [15] argued that $R_S$ can be estimated from the differences between maximum and minimum air temperatures as follows:

$$R_S = a \cdot (T_{\max} - T_{\min})^b R_a \tag{1}$$

where $R_a$ is the extraterrestrial radiation (MJ m$^{-2}$ d$^{-1}$), $T_{\max}$ and $T_{\min}$ are maximum and minimum air temperatures (°C) and $a$, $b$ are empirical coefficients. $b$ is usually set to 0.5, and $a$ differs in interior ($a$ = 0.16) and coastal ($a$ = 0.19) regions. In this study a value of $a$ = 0.16 was used.

According to Mahmood and Hubbard [16] the atmospheric transmissivity in a given day is a function of day of year (*DOY*). Daily incoming global solar radiation ($R_S$) is correlated to daily range of temperature and extraterrestrial radiation as follows:

$$R_S = c \cdot (T_{\max} - T_{\min})^d ICSKY^e \tag{2}$$

where $c$ = 0.182, $d$ = 0.69, $e$ = 0.91 and *ICSKY* is the corrected clear-sky solar irradiation and can be computed with the following equation:

$$ICSKY = I_S T \tag{3}$$

In which the term $T$ represents the empirical transmissivity coefficients:

$$T = 0.8 + 0.12\left(\frac{|182 - DOY|}{183}\right)^{15} \tag{4}$$

And the term $I_S$ denotes the clear day solar-irradiation (MJ m$^{-2}$ d$^{-1}$):

$$I_S = 0.04188\left(A + B\sin\left[(DOY + 10.5)\frac{2\pi}{365}\frac{\pi}{2}\right]\right) \tag{5}$$

In Eq. (5) the $A$ and $B$ coefficients are defined as follows [17]:

$$A = \left\{\sin\phi(46.355LD - 574.3885) + 816.41\cos\phi\sin\left[\left(LD\frac{\pi}{24}\right)\right]\right\}(0.29\cos\phi + 0.52) \tag{6}$$

$$B = \left\{\sin\phi(574.3885 - 1.509LD) - 29.59\cos\phi\sin\left[\left(LD\frac{\pi}{24}\right)\right]\right\}(0.29\cos\phi + 0.52) \tag{7}$$

And the term *LD* (longest *DOY*) is estimated as follows [17]:

$$LD = 0.267\sin^{-1}\left[0.5 + \left(\frac{0.007895}{\cos\varphi}\right) + (0.2168875\tan\varphi)\right]^{0.5} \tag{8}$$

Mahmood and Hubbard [16] proposed an approach to locally adjust the MH1 model in order to account for advection and frontal movements in local scales. The proposed approach includes the following linear regression model (so-called MH2):

$$R_S = \frac{R_{biased} - f}{g} \tag{9}$$

where $R_{biased}$ is the MH1 model result, $f$ equals 2.499 MJ m$^{-2}$ d$^{-1}$ and $g$ equals 0.802.

### 2.2. Artificial Neural Networks

Artificial Neural Networks (ANNs) are parallel information-processing systems. Artificial Neural Networks (ANNs) were originally designed for modeling the performance of a biological neural system. The internal structure of ANNs is similar to the structure of a biological brain with a number of layers of fully interconnected nodes or neurons. Each neuron is connected to other neurons by means of direct communication links, each one with an associated weight. The most common architecture is composed of: the input layer, where the data are introduced into the ANN, the hidden layer(s) where the data are processed, and the output layer where the results of given inputs are obtained. This type of ANN is called multilayer perceptron (MLP) [18]. Each neuron receives a transformed linear combination of the outputs of the previous neurons ($\sum w_{ij}x_j$). This linear combination is characterized by some parameters applied to each output, which are the weights or numerical estimates of the connection strength between the neurons ($w_{ij}$). An activation threshold is also assigned to each neuron ($w_0$). This activation threshold is analogous to an independent term of the linear combination of the outputs from the previous neurons, and it is considered as a weight assigned to a fictitious neuron with an output value of 1. Then, an activation non-linear function is applied to the linear combination of outputs ($f(\sum w_{ij}x_j)$). This activation function, which is usually a sigmoid function, limits the values of the output of each neuron to values between two asymptotes. After the application of the activation function, the output of each neuron goes to those of the next layer. As a result of the application of an ANN, a model characterized by a very complex and flexible equation is obtained, which contains a large number of parameters (weights of the links between the neurons, included the activation threshold), giving a very high capacity of approximation to the final output equation. These parameters (weights that define the connections between the neurons) are adjusted ($\Delta w_{ji}$) by using known inputs and outputs in an iterative process called neural network training. The objective of this training is to minimize the error function ($\xi$), which is constructed from the discrepancy between the predicted and expected values. The detailed theoretical information about ANNs can be found in Bishop [19] or Haykin [20]. In Section 3.3 there is a description of the ANN procedure adopted in this study.

### 2.3. Adaptive Neuro-Fuzzy Inference System

The Adaptive Neuro-Fuzzy Inference System (ANFIS) was firstly introduced by Jang [21] and later on widely applied in various problems. ANFIS is a combination of an adaptive neural network (ANN) and a fuzzy inference system. The parameters of the fuzzy inference system are determined by the neural network learning algorithms. ANFIS is capable of approximating any real continuous function on a compact set of parameters to any degree of accuracy

[22]. ANFIS identifies a set of parameters through a hybrid learning rule combining back propagation gradient descent error digestion and a least squared error method. There are mainly two approaches for fuzzy inference systems, namely the approaches of Mamdani [23] and Sugeno [24]. The differences between the two approaches arise from the consequent part where Mamdani's approach uses fuzzy membership functions, while linear or constant functions are used in Sugeno's approach. In this study the Sugeno method was applied for modeling the incoming global solar radiation. The detailed theoretical information about ANFIS can be found in the papers mentioned above. In Section 3.4 there is description of the Sugeno procedure adopted in this study.

### 2.4. Genetic Programming

The methodology of Genetic Programming (GP) was first proposed by Koza [25], as a generalization of Genetic Algorithms (GAs) [26]. GAs are combinatorial optimization methods that search for solution using an analogy between optimization and natural selection. The methodology of GAs involves coding, fitness function computation, and operations of reproduction, crossover and mutation of individuals [26]. The fundamental difference between GP and GAs lies in the nature of individuals, where in GAs individuals are linear strings of fixed length (as chromosomes), while in GP individuals are non-linear entities of different sizes and shapes (as parse trees). The GP algorithms firstly define an objective function as a qualitative criterion. Next, this function is used for measurement and evaluation of different solutions in a step by step manner of structural correction until GP leads to a suitable solution. GP is an evolutionary algorithm (EA) and is popular because of its high accuracy. Major advantages of GP are that it can be applied to areas where: (i) the interrelationships among the relevant variables are poorly understood (or where it is suspected that the current understanding may well be less than satisfactory); (ii) finding the ultimate solution is hard; (iii) conventional mathematical analysis does not, or cannot, provide analytical solutions; (iv) an approximate solution is acceptable (or is the only result that is ever likely to be obtained); (v) small improvements in the performance are routinely measured (or easily measurable) and highly valued, and; (vi) there is a large amount of data, in computer readable form, that requires examination, classification, and integration (such as satellite observations) [27]. The applied Genetic Programming variant in the present study is Gene Expression Programming (GEP). In GEP, the individuals are non-linear structures of different size and shape (expression trees) that are encoded by linear chromosomes composed of multiple genes, each gene encoding a smaller subprogram. Thus, in GEP, the genotype (the linear chromosomes) and the phenotype (the expression trees) are different entities (both structurally and functionally). In GEP all the genetic modifications take place in the linear chromosomes, so only the linear chromosomes are transmitted in the process of reproduction. Furthermore, the structural and functional organization of the chromosomes allows the unconstrained operation of important genetic operators such as mutation, transposition and recombination. And as in nature, it's only during development that the information encoded in the chromosomes is finally expressed into expression trees. One strength of the GEP approach is that the creation of genetic diversity is extremely simplified as genetic operators work at the chromosomes level. Another strength of GEP consists of its unique, multigenic nature which allows the evolution of more complex programs composed of several subprograms. As a result GEP surpasses the old GP system in 100–10,000 times [28]. The advantages of a system like GEP are clear from nature, but the most important are [28]: (i) the chromosomes are simple entities: linear, compact, relatively small, easy to manipulate genetically (replicate, mutate, recombine, etc.); (ii) the expression trees are exclusively the expression of their respective chromosomes, they are entities upon which selection acts, and according to fitness, they are selected to reproduce with modification. The detailed theoretical information about GEPs can be found in the papers mentioned above. In Section 3.5 there is a description of the GEP procedure adopted in this study.

## 3. Implementation of the models

### 3.1. Used data

Climatic data from four weather stations in the Basque region of Alava situated in Northern Spain, covering a period of 5 years (1999–2003) were analyzed for estimating the incoming global solar radiation. The region of Alava (between the parallels 43° and 42°30′) is in the south of the Basque Country with 3037 km$^2$. Agricultural activity in the Basque Country is mainly concentrated in Alava with 123,000 ha of agricultural land. The climate of this region is a transition climate between the Atlantic climate of the north of the region and the continental Mediterranean of the south. The annual mean temperature of the region is about 11–12 °C and the annual mean precipitation ranges between 600 and 800 mm.

Data from four weather stations, namely, Arkaute, Salvatierra, Navarrete and Zambrana were used in this study. Fig. 1 displays the geographical position of these weather stations. The aforementioned stations are equipped with electronic sensors for the detection of the air temperatures (R018, Rotronic), and global solar radiation (K614, Kipp&Zonne). These sensors provide meteorological observations each 10 min. The applied data sample consisted of daily maximum and minimum air temperature and recorded daily global solar radiation values.

The input parameters of the artificial intelligence models as well as empirical equations were different combinations of daily values of extraterrestrial radiation ($R_a$), maximum air temperature ($T_{max}$), minimum air temperature ($T_{min}$), day of year (*DOY*) and corrected clear-sky solar irradiation (*ICSKY*). Daily incoming global solar radiation ($R_S$) was the applied models' output. Thus, the following input combinations were studied in this paper:

(i) $R_a$,
(ii) $R_a$, $T_{max}$,
(iii) $R_a$, $T_{max}$, $T_{min}$,
(iv) $R_a$, $T_{max}$, $T_{min}$, *DOY*,
(v) $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY*.

The description of the different weather stations and the mean weather data from each station for the period of study (1999–2003) are given in Table 1. There are few differences among the stations and not any significant local microclimatic effects can be distinguished among them. For this reason the training and optimization of the models was based on a pool approach in order to develop regional models of estimation of $R_S$.

The period from 1999 to 2001 of daily meteorological observations was used as training and optimization data (pool of the four stations, 4220 observations). In the case of artificial intelligence models, and in order to avoid over-learning the available training data (1999–2001) were split into two subsets: 75% of the observations for training and 25% for cross validation. The period from 2002 to 2003 was used for an independent validation of the models (pool of the four stations, 2855 observations).

### 3.2. Empirical Equations Optimization

The HS, MH1, and MH2 equations presented in the previous section were compared with the artificial intelligence models

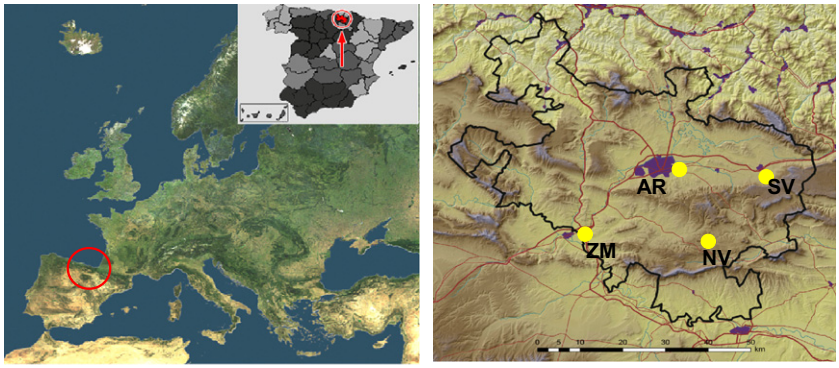


**Fig. 1.** Map of situation of the Basque Meteorological Service's weather stations used in this study (AR: Arkaute. SV: Salvatierra. NV: Navarrete. ZM: Zambrana).

**Table 1**
Summary of the weather station sites used in this study (means of daily data for the period 1999–2003).

| Station | UTM coordinates | | | $N$ | Meteorological parameters | | |
|---|---|---|---|---|---|---|---|
| | Longitude (°) | Latitude (°) | Height (m) | | $T_{mean}$ (°C) | $\Delta T$ (°C) | $R_s$ (MJ m$^{-2}$ d$^{-1}$) |
| Arkaute | 2.63 | 42.85 | 517 | 1766 | 11.22 | 11.12 | 12.33 |
| Salvatierra | 2.39 | 42.86 | 589 | 1789 | 11.59 | 10.77 | 12.34 |
| Navarrete | 2.52 | 42.64 | 689 | 1826 | 9.90 | 10.95 | 12.17 |
| Zambrana | 2.89 | 42.67 | 470 | 1823 | 11.13 | 12.31 | 12.62 |

*Note*: $\Delta T$, Difference between maximum and minimum temperature; $R_s$, solar radiation; $N$, number of patterns after the elimination of inaccurate data; $T_{mean}$, daily mean temperature.

following two approaches: using the original equations presented in Section 2.1; and using an optimized version of these equations. The optimization of these empirical equations was based on the utilization of a standard unconstrained non-linear scheme. This entails to find the local minimum of a continuously differentiable real-valued function of n variables from a given start point $x_0$ [10]. To achieve this, Levenberg–Marquardt algorithm was used, which is the default algorithm for unconstrained models of the SPSS 17.0 software package [29]. The $a$ and $b$ parameters of the HS model (Eq. (1)) were optimized across the training dataset, and the performance of the locally optimized HS model was evaluated through the validation dataset. The $a$ (0.16) and $b$ (0.5) values proposed by Hargreaves and Samani [15] were used as initial estimates. The same procedure was applied to the optimization of the MH1 model (Eq. (2)), using the $c$ (0.182), $d$ (0.690) and $e$ (0.910) values of Mahmood and Hubbard [16] for the initialization. Finally, the MH2 model (Eq. (9)) was optimized using the initial $f$ (2.499) and $g$ (0.802) values of Mahmood and Hubbard [16].

### 3.3. Artificial Neural Networks (ANNs) modeling of $R_S$

Daily observed global solar radiation values were used as training data for the implementation of ANNs for each one of the five input combinations presented above. The evaluation of their usefulness for global solar radiation estimation was based on the comparison of their performance with the empirical equations. The selection of the ANNs architectures was based on the employment of selection algorithms integrated in the IPS (Intelligent Problem Solver) module of Statistica Neural Networks software [30]. The IPS module replaces the traditional heuristic process of an ANN design by interleaved search algorithms that determine the selection of inputs (if necessary), the number of hidden units, and other key factors in the network design. The IPS module works by experimentally sampling a range of network configurations. It automatically determines the network complexity, using a variety of algorithms for different network types. Nevertheless in the case of the present study the number of hidden units was determined manually selecting the advance version of the IPS module. Multilayer Perceptron ANNs with a number of hidden units from 1 to 12 were generated for each combination of inputs. The ANNs with the best performance for each combination of inputs were selected. Prior to the application of the IPS module, data were divided into the different groups mentioned in the previous subsection for the implementation and validation of the Artificial Neural Network models.

The global outputs of each ANN were compared with the training values, or expected $R_S$ outputs, for the construction of a multidimensional error function, which is a function of the weights defining the links between the neurons. This iterative training process was automatically conducted by the IPS module using four different minimization algorithms: back propagation, Levenverg–Marquardt, conjugate gradient descent and quick propagation. The IPS module was executed selecting the "thorough" option which gives the IPS module the opportunity to deploy detailed algorithms. After the application of the IPS module, a non-automatically supervised training was applied to the model selected by the IPS module in order to improve its performance. The quick propagation algorithm (Eq. (10)) was used in this non-automatic training with a maximum of 50,000 iterations. The ANN with the best performance for each combination of inputs was retained and selected. In the case of detecting over-learning or convergence to a local minimum, the training process was stopped.

$$\Delta \mathbf{w}(s) = \frac{\nabla \xi|_{\mathbf{w}(s)}}{\nabla \xi|_{\mathbf{w}(s-1)} - \nabla \xi|_{\mathbf{w}(s)}} \Delta \mathbf{w}(s-1) \tag{10}$$

where $\xi^n$ is the error function, $\mathbf{w}$ is the vector of weights (internal parameters of the ANN), and $s$ is the iteration number. In Fig. 2 there is a description of the ANN with the best performance after the training process, the four input ANN ($R_a$, $T_{max}$, $T_{min}$, $DOY$). Fig. 2 also contains a general description of the structure of a hidden neuron, according to the explanation of Section 2.2.

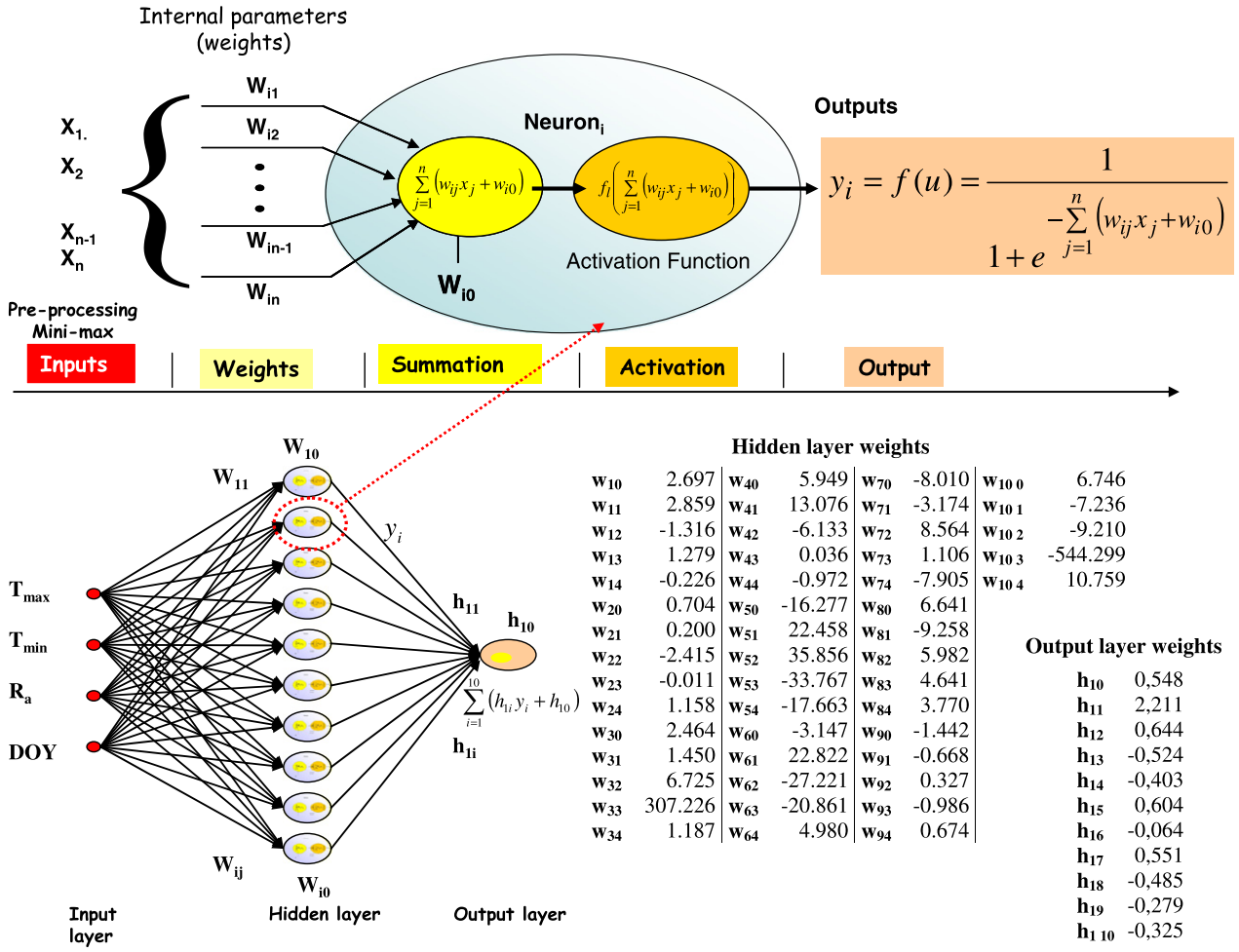


**Fig. 2.** Description of the ANN with the best performance after the training process, the four input ANN ($R_a$, $T_{max}$, $T_{min}$, $DOY$).

### 3.4. Modeling daily $R_S$ by using Adaptive Neuro-Fuzzy Inference System (ANFIS)

The procedure described by Jang et al. [22] was adopted for modeling daily $R_S$. This procedure is functionally equivalent to the Sugeno first-order fuzzy model. As a simple example of the procedure adopted, a fuzzy inference system with two inputs $x$ and $y$ and one output $f$ is assumed. Here, $x$ and $y$ might be considered as extraterrestrial radiation ($R_a$) and maximum air temperature ($T_{max}$), respectively, while the output $f$ would represent the incoming global solar radiation ($R_S$). This is an example with one of the input combinations used in this study.

Suppose that the rule base contains two fuzzy IF–THEN rules:

$$\text{Rule 1}: \text{ IF } x \text{ is } A_1 \text{ and } y \text{is } B_1, \text{THEN } f_1 = p_1 x + q_1 y + r_1 \quad (11)$$

$$\text{Rule 2}: \text{ IF } x \text{ is } A_2 \text{ and } y \text{is } B_2, \text{THEN } f_2 = p_2 x + q_2 y + r_2 \quad (12)$$

The IF (antecedent) part is fuzzy in nature, while the THEN (consequent) part is a crisp function of an antecedent variable (as a rule, a linear equation). The study presented here for modeling global solar radiation, for the above example Eqs. (1) and (2) can be written as:

$$\text{Rule1}: \text{IF } R_a \text{ is LOW and } T_{max} \text{ is LOW, THEN } R_S = p_1 R_a + q_1 T_{max} + r_1$$

$$\text{Rule2}: \text{IF } R_a \text{ is HIGH and } T_{max} \text{ is MEDIUM, THEN } R_S = p_2 R_a + q_2 T_{max} + r_2$$

.....

where $p_i$, $q_i$ and $r_i$ are parameters with $i = 1, 2, 3, \ldots, n$ corresponding to Rule 1, Rule 2, Rule 3, ..., Rule $n$. The resulting Type 3 Sugeno fuzzy reasoning model is shown in Fig. 3 (General ANFIS structure, on the left of the figure). In a Type 3 Sugeno fuzzy model, the output of each rule is a linear combination of input variables plus a constant term and the final output $f$ is a weighted average of each rule output. The corresponding equivalent ANFIS architecture is represented in Fig. 3 (General ANFIS structure, on the left of the figure). The node function in the same layer of the same function family is described as follows [21]:

*Layer 1:* Every node $i$ in this layer is an adaptive node with node function.

$$O_i^1 = \mu Ai(R_a) \quad (13)$$

where $R_a$ is the input to the $i$th node and $A_i$ is a linguistic label (such as HIGH or LOW) associated with this node function. A similar equation as Eq. (12) may be considered for the input $T_{max}$.

The node function $O_i^l$ is the membership function of $A_i$ and specifies the degree to which the given input $R_a$ (or $T_{max}$) satisfies the quantifier $A_i$. The membership function for $A$ is usually described by bell-functions, e.g.

$$A_i(R_a) = \frac{1}{1 + [(R_a - c_i)/a_i]^{2b_i}} \quad (14)$$

where $\{a_i, b_i, c_i\}$ is the parameter set and $\mu$ is the membership function of $A_i$. As the values of these parameters change, the bell-shaped function varies accordingly, thus exhibiting various forms of membership functions depending on the linguistic label $A_i$. In fact, any

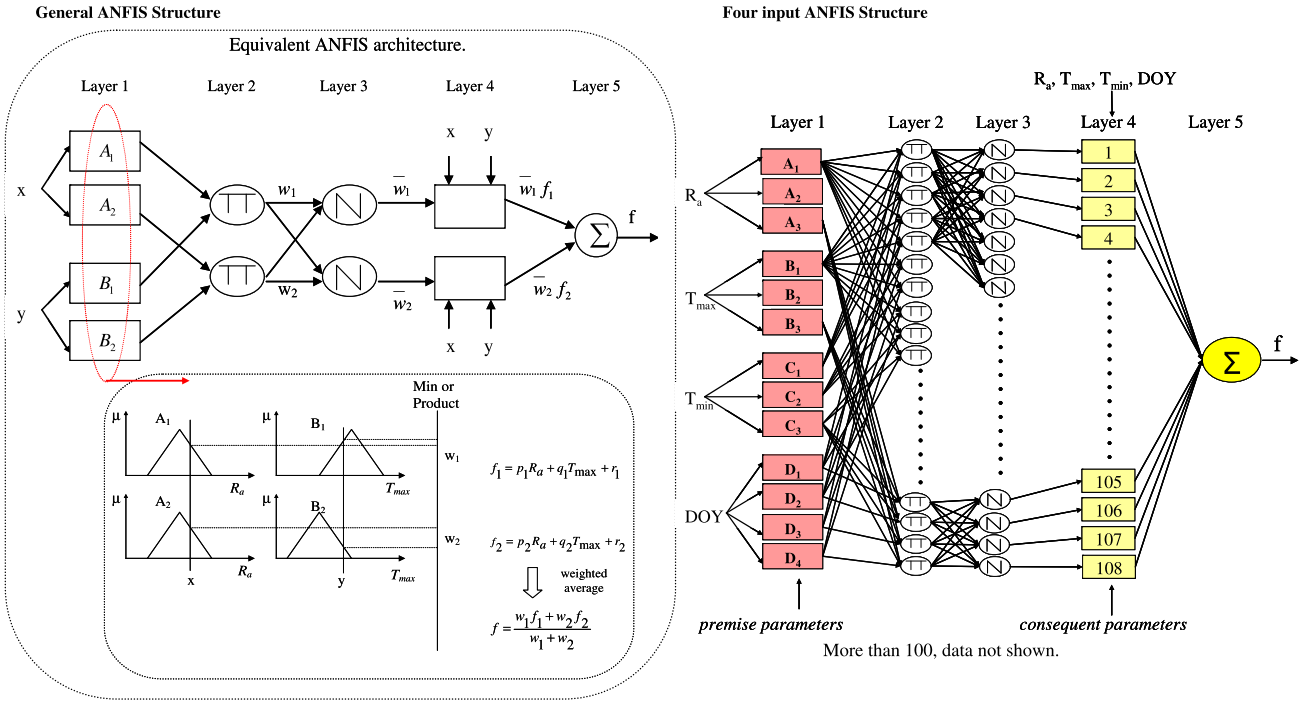


**Fig. 3.** General ANFIS structure and four input ANFIS structure (the ANFIS with the best performance after the training process).

continuous and piecewise differentiable functions, such as commonly used triangular or trapezoidal membership functions, are also qualified candidates for node function in this layer. Parameters in this layer are referred to *premise parameters*.

*Layer 2:* This layer consists of circle nodes labeled $\Pi$ which multiplies the incoming signals and sends the product out. For instance

$$O_i^2 = w_i = A_i(R_a)B_i(T_{\max}), \quad i = 1,2 \tag{15}$$

Each node output represents the firing strength of a rule.

*Layer 3:* In this layer, the circle nodes labeled *N*, calculate the ratio of the *i*th rule firing strength to the sum of all rule firing strengths

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2}, \quad \text{for } i = 1,2 \tag{16}$$

The outputs of this layer are referred to as *normalized firing strengths*.

*Layer 4:* All of the nodes in this layer are adaptive with a node function.

$$O_i^4 = \bar{w}_i \cdot f_i = \bar{w}_i(p_i R_a + q_i T_{\max} + r_i) \tag{17}$$

where $\bar{w}_i$ is the output of layer 3, and $\{p_i, q_i, r_i\}$ is the parameter set. Parameters in this layer are called *consequent parameters*.

*Layer 5:* The single circle node of this layer, labeled $\Sigma$, computes the overall outputs as the summation of all incoming signals.

$$O_i^5 = \sum \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{18}$$

Thus an adaptive network has been constructed, which is functionally equivalent to a Type 3 fuzzy inference system.

ANFIS learning employs two methods for updating membership function parameters: (i) back propagation for all parameters (a steepest descent method); and (ii) a hybrid method consisting of back propagation for the parameters associated with the input membership (*premise parameters*) and least squares estimation for the parameters associated with the output membership functions (*consequent parameters*). The hybrid training method was applied for modeling $R_S$. Although the ANN parameters are provided in Fig. 2, it is not possible to provide the ANFIS parameters because the parameters are too much. Two different program codes, including fuzzy logic and ANN toolboxes, were written in MATLAB language for the ANFIS simulations. In Fig. 3 there is a description of the ANFIS with the best performance after the training process, the four-input ANFIS ($R_a$, $T_{\max}$, $T_{\min}$, *DOY*).

### 3.5. Derivation of $R_S$ model based on the application of Gene Expression Programming (GEP)

GeneXproTools software package [31] was used for the implementation of GEP models. The procedure to model daily incoming global solar radiation is as follows. The first step is the selection of fitness function. For this problem, the fitness function, $f_i$, of an individual program, *i*, is expressed as [23]: $f_i = \sum_{j=1}^{n}(M - |C_{i,j} - T_j|)$; in which *M* is the range of selection, $C_{i,j}$ is the value predicted by individual program *i* for fitness case *j*, and $T_j$ is the target value for fitness case *j*. For a perfect fit, $C_{i,j} = T_j$. The second step consists of choosing the set of terminals *T* and the set of functions *F* to create the chromosomes. In the current problem, the terminal set includes the following variables: {$R_a$, $T_{\max}$, $T_{\min}$, *DOY* and *ICSKY*}. The choice of the appropriate function depends on the viewpoint of user. In this study, different mathematical functions were utilized $\left(\text{i.e.}, \{+,-,*,/\}, \left\{\sqrt{}, \sqrt[3]{}, \ln(x), e^x, x^2, x^3\right\}\right)$. The third step is to choose the chromosomal architecture. Length of head, $h = 8$, and three genes per chromosomes were employed. The fourth step is to choose the linking function, which was "addition" for this study. The linking function must be chosen as "addition" or "multiplication" for algebraic sub trees [28]. Here, the sub trees (ET) are linked by addition. The final step is to select the GEP operators. The learning algorithms of GEP apply the following basic operators: mutation (allows the evolution of good solutions for the studied models to virtually all problems); inversion (which is restricted to the heads of genes), one-point recombination (in which the parent chromosomes are paired and split up at exactly the same

point), two-point recombination (in which two parent chromosomes are paired and two points are randomly chosen as crossover points), gene recombination (in which entire genes are exchanged between two parent chromosomes, forming two daughter chromosomes containing genes from both parents), gene transposition (in which an entire gene works as a transposon and transposes itself to the beginning of the chromosome), IS transposition (short fragments of the genome with a function or terminal in the first position that transpose to the heads of gene except the root) and RIS transposition (short fragments with a function in the first position that transpose to the start position of genes) [28]. In Fig. 4 there is a description of the GEP implementation procedure described above, and of the GEP model with the best performance after the training process, the five-input GEP ($R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY*).

## 4. Results and discussion

### 4.1. Performance evaluation criteria

Five statistical evaluation criteria were used to assess the models' performances:

(i) Root mean square error (RMSE) defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{Sio} - R_{Sie})^2} \tag{19}$$

(ii) Mean absolute error (MAE) defined as:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|R_{Sio} - R_{Sie}|}{n} \tag{20}$$

where $R_{Sio}$ and $R_{Sie}$ denote the observed and estimated incoming global solar radiation values.

The scale dependency of the RMSE and MAE indices can be overcome by using a skill score, which compares the performance of a specified simulation with the performance of reference simulation. The RMSE- and MAE-based skill scores can be computed as follows:

(iii) RMSE-based skill score ($SS_{RMSE}$):

$$SS_{RMSE} = \left(1 - \frac{\mathrm{RMSE}}{\sqrt{(1/n)\sum_{i=1}^{n}(R_{Sio} - \overline{R})^2}}\right) \tag{21}$$

(iv) MAE-based skill score ($SS_{MAE}$):

$$SS_{MAE} = \left(1 - \frac{\mathrm{MAE}}{\sqrt{(1/n)\sum_{i=1}^{n}(R_{Sio} - \overline{R})^2}}\right) \tag{22}$$

These statistical indexes were used by Fortin et al. [10] in their study. A *SS* of 1 represents a perfect fit between predicted and observed values, and 0 indicates that performance of the model is no better than a "no-knowledge" model giving the mean of observations at every time step.

(v) The $r^2$ criterion of Nash and Sutcliffe [32]:

$$r^2 = \frac{SS - SS_{ref}}{1 - SS_{ref}} \tag{23}$$

where the $SS_{ref}$ denotes the skill score value of HS model. This criterion expresses the proportion of initial variance unaccounted for by a reference model (the HS model in our case) that may subsequently be accounted for by other models. Negative values of this index indicate that the alternative model has negative effects on the performance [10].

### 4.2. Empirical equation results

The performance of the different models evaluated in this study is presented in Table 2. The Hargreaves–Samani model (HS) has been used as the reference model for $r^2$ calculation due to its simplicity. The results are divided into four groups: empirical equations (non-optimized and optimized), ANNs, ANFIS, and GEPs. The results about the last three groups (artificial intelligence models) will be discussed later.

Regarding the first group, the Hargreaves and Samani [15] equation (HS) presents values of 0.58 and 0.52 for the MAE-based skill score ($SS_{MAE}$), and the RMSE skill score ($SS_{RMSE}$) respectively for the validation period (RMSE value of 3.94 MJ m$^{-2}$ d$^{-1}$). So according the interpretation of these statistical indexes ($SS_{MAE}$ and $SS_{RMSE}$) mentioned above the model performs better than a "no-knowledge" model that would always give as prediction the average value of the phenomenon. The performance of the HS model in the conditions of this study is better than the obtained by Fortin et al. [10] in Canada, and similar to the obtained by Liu et al. [33] in China. The empirical equations of Mahmood and Hubbard [16] (MH1 and MH2) (non-optimized) improve clearly the performance of the HS model with RMSE values of 3.21 MJ m$^{-2}$ d$^{-1}$ (MH1) and 3.38 MJ m$^{-2}$ d$^{-1}$ (MH2) for the validation period, and $r^2$ values of 18.64 ($r^2_{MAE}$) and 14.27 ($r^2_{RMSE}$). As it was mentioned above these $r^2$ values express the proportion of initial variance unaccounted for by the HS model that may be accounted for by MH1 and MH2 models. The performance of these models, MH1 and MH2 (non-optimized version), is significantly better than the results obtained by Fortin et al. [10] in the conditions of Canada (Montreal), and similar to those obtained by Mahmood and Hubbard [16] in North Dakota and South Dakota (USA).

As it is shown in Fig. 5 HS equation is more biased than the MH1 and MH2 equations, and tends to overestimate $R_S$, moreover for $R_S$ low values. This could explain the differences of performances between the HS equation and the Mahmood and Hubbard equations.

HS, MH1 and MH2 equations were optimized using the period 1999–2001 following a procedure based on a standard unconstrained non-linear scheme (see Section 3.2). The HS model optimization procedure leaded to parameter values of: $a = 0.094$ and $b = 0.636$ ($HS^{op}$). As a result of this optimization procedure the performance of the HS model was significantly improved (MAE based skill score ($r^2_{MAE} = 21.8\%$), RMSE based skill score ($r^2_{MAE} = 20.4\%$), and a RMSE value of 3.14 MJ m$^{-2}$ d$^{-1}$). So it seems advisable to replace the original HS equation by the optimized one ($HS^{op}$). The MH1 and MH2 optimization procedures leaded to parameter values of: $c = 0.085$, $d = 0.619$, $e = 1.225$ ($MH1^{op}$); and $f = 1.002$, $g = 0.948$ ($MH2^{op}$). In this case there was not a very important improvement of the performance of these models compared with their non-optimized versions which present a satisfactory performance in the area of study (Table 2). According to Table 2 the optimized HS equation presents the best performance of all the studied empirical equations, so this is the recommended model among the empirical equations. Nevertheless the original MH1 equation (none optimized) is an interesting alternative.

### 4.3. Artificial intelligence models' results

The total performances of the artificial intelligence models for the training and validation period (1999–2003) by using the whole data set of the all four studied stations are given in Table 2.

Table 2 also represents the final architectures of ANFIS and ANN models. The second column represents the combination of inputs

**Table 2**
Statistical summary of the implemented and applied models during the study period.

| Model | Inputs | Structure | Training/optimization period (1999–2001) | | | | | | Validation period (2002–2003) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MAE ($MJ\ m^{-2}\ d^{-1}$) | $SS_{MAE}$ ($MJ\ m^{-2}\ d^{-1}$) | $r^2_{MAE}$ (%) | RMSE ($MJ\ m^{-2}\ d^{-1}$) | $SS_{RMSE}$ ($MJ\ m^{-2}\ d^{-1}$) | $r^2_{RMSE}$ (%) | MAE ($MJ\ m^{-2}\ d^{-1}$) | $SS_{MAE}$ ($MJ\ m^{-2}\ d^{-1}$) | $r^2_{MAE}$ (%) | RMSE ($MJ\ m^{-2}\ d^{-1}$) | $SS_{RMSE}$ ($MJ\ m^{-2}\ d^{-1}$) | $r^2_{RMSE}$ (%) |
| *Empirical equations* | | | | | | | | | | | | | | |
| HS | $R_a$, $T_{max}$, $T_{min}$ | Non-calibrated | 3.25 | 0.53 | 0.00 | 4.22 | 0.49 | 0.00 | 3.02 | 0.58 | 0.00 | 3.94 | 0.52 | 0.00 |
| MH1 | $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY* | Non-calibrated | 2.60 | 0.63 | 20.11 | 3.41 | 0.58 | 19.14 | 2.43 | 0.66 | 19.78 | 3.21 | 0.61 | 18.64 |
| MH2 | $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY* | Non-calibrated | 2.73 | 0.61 | 16.15 | 3.58 | 0.56 | 15.18 | 2.55 | 0.64 | 15.55 | 3.38 | 0.59 | 14.27 |
| $HS^{op}$ | $R_a$, $T_{max}$, $T_{min}$ | Non-linear optimization | 2.47 | 0.65 | 24.14 | 3.29 | 0.60 | 21.92 | 2.36 | 0.67 | 21.82 | 3.14 | 0.62 | 20.43 |
| $MH1^{op}$ | $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY* | Non-linear optimization | 2.46 | 0.65 | 24.28 | 3.28 | 0.60 | 22.13 | 2.39 | 0.67 | 20.79 | 3.16 | 0.62 | 19.88 |
| $MH2^{op}$ | $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY* | Non-linear optimization | 2.56 | 0.63 | 21.35 | 3.44 | 0.58 | 18.36 | 2.43 | 0.66 | 19.52 | 3.29 | 0.60 | 16.47 |
| *Artificial Neural Networks* | | | | | | | | | | | | | | |
| ANN1 | $R_a$ | MLP (1–7–1) | 3.94 | 0.43 | −20.95 | 5.03 | 0.39 | −19.20 | 4.07 | 0.43 | −34.64 | 5.18 | 0.37 | −31.34 |
| ANN2 | $R_a$, $T_{max}$ | MLP (2–6–1) | 2.97 | 0.57 | 8.86 | 3.86 | 0.53 | 8.51 | 3.14 | 0.56 | −4.00 | 4.04 | 0.51 | −2.40 |
| ANN3 | $R_a$, $T_{max}$, $T_{min}$ | MLP (3–6–1) | 2.36 | 0.66 | 27.61 | 3.16 | 0.61 | 24.98 | 2.24 | 0.69 | 25.82 | 2.97 | 0.64 | 24.77 |
| ANN4 | $R_a$, $T_{max}$, $T_{min}$, *DOY* | MLP (4–10–1) | 2.31 | 0.67 | 28.93 | 3.11 | 0.62 | 26.18 | 2.22 | 0.69 | 26.69 | 2.93 | 0.64 | 25.60 |
| ANN5 | $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY* | MLP (5–12–1) | 2.35 | 0.66 | 27.85 | 3.15 | 0.62 | 25.37 | 2.23 | 0.69 | 26.36 | 2.93 | 0.65 | 25.63 |
| *Adaptative Neuro-Fuzzy Inference Systems* | | | | | | | | | | | | | | |
| ANFIS1 | $R_a$ | 3 | 3.95 | 0.43 | −21.37 | 5.05 | 0.38 | −19.71 | 4.08 | 0.43 | −34.93 | 5.19 | 0.37 | −31.70 |
| ANFIS2 | $R_a$, $T_{max}$ | 4 and 3 | 2.95 | 0.58 | 9.31 | 3.84 | 0.53 | 8.84 | 3.19 | 0.56 | −5.38 | 4.09 | 0.50 | −3.73 |
| ANFIS3 | $R_a$, $T_{max}$, $T_{min}$ | 4, 3 and 3 | 2.65 | 0.62 | 18.51 | 3.49 | 0.57 | 17.21 | 2.55 | 0.64 | 15.72 | 3.33 | 0.60 | 15.58 |
| ANFIS4 | $R_a$, $T_{max}$, $T_{min}$, *DOY* | 3, 3, 3 and 4 | 2.31 | 0.67 | 29.11 | 3.09 | 0.62 | 26.63 | 2.35 | 0.67 | 22.21 | 3.14 | 0.62 | 20.40 |
| ANFIS5 | $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY* | 3, 3, 3, 3 and 3 | 2.27 | 0.67 | 30.28 | 3.04 | 0.63 | 27.91 | 2.39 | 0.67 | 21.03 | 3.14 | 0.62 | 20.26 |
| *Gene Expression Programming systems* | | | | | | | | | | | | | | |
| GEP1 | $R_a$ | See Table 3 | 3.95 | 0.43 | −21.35 | 5.05 | 0.38 | −19.77 | 4.09 | 0.43 | −35.34 | 5.20 | 0.37 | −31.84 |
| GEP2 | $R_a$, $T_{max}$ | See Table 3 | 3.11 | 0.55 | 4.36 | 4.00 | 0.51 | 5.03 | 3.24 | 0.55 | −7.15 | 4.12 | 0.50 | −4.57 |
| GEP3 | $R_a$, $T_{max}$, $T_{min}$ | See Table 3 | 2.65 | 0.62 | 18.51 | 3.49 | 0.57 | 17.21 | 2.55 | 0.64 | 15.72 | 3.33 | 0.60 | 15.58 |
| GEP4 | $R_a$, $T_{max}$, $T_{min}$, *DOY* | See Table 3 | 2.75 | 0.61 | 15.58 | 3.59 | 0.56 | 14.83 | 2.68 | 0.63 | 11.46 | 3.49 | 0.58 | 11.57 |
| GEP5 | $R_a$, $T_{max}$, $T_{min}$, *DOY*, *ICSKY* | See Table 3 | 2.64 | 0.62 | 18.78 | 3.49 | 0.57 | 17.25 | 2.53 | 0.65 | 16.38 | 3.31 | 0.60 | 16.01 |

*Note*: $R_a$, Extraterrestrial radiation; $T_{mean}$, daily mean temperature; $T_{max}$, daily maximum temperature; $T_{min}$, daily minimum temperature; *DOY*, day of the year; *ICSKY*, corrected clear-sky solar irradiation.

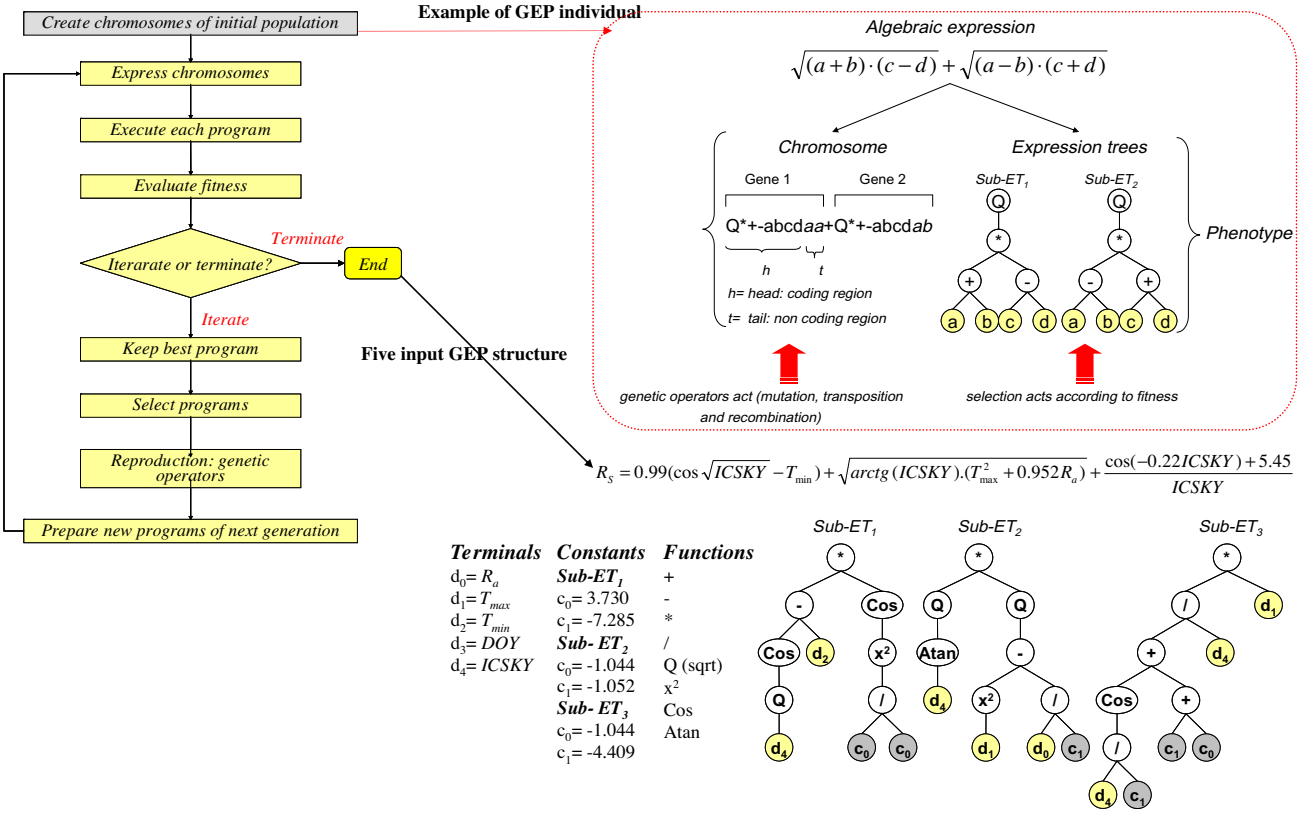


**Fig. 4.** General GEP structure and five input GEP structure (the GEP with the best performance after the training process).

of each AI model. The following input combinations have been evaluated in this study:

(i) $R_a$,
(ii) $R_a$, $T_{max}$,
(iii) $R_a$, $T_{max}$, $T_{min}$,
(iv) $R_a$, $T_{max}$, $T_{min}$, *DOY*,
(v) $R_a$, $T_{max}$, $T_{min}$, *DOY, ICSKY*.

The third column of Table 2 indicates the number of membership functions of each input variable of ANFIS models and the number of input, hidden and output nodes of each ANN model. For instance, in the input combination (iv), the ANFIS model (ANFIS4) has 3, 3, 3 and 4 triangular membership functions for the input variables, $R_a$, $T_{max}$, $T_{min}$, and *DOY*, respectively (Fig. 3). It should be noted that fuzzy membership functions can take many forms, but triangular membership functions are often selected for practical applications [34]. Similarly, for input combination (iv), the quadruple-input ANN model has 4, 10 and 1 neurons for the input, hidden and output layers, respectively. The number of the input layer nodes is corresponded to the applied input variables where as the number of the output layer nodes corresponded to the output variable (i.e. global solar radiation). Table 3 represents the final formulation of each GEP model.

According to Table 2, the optimal results for GEP models are obtained with input combination (iii) (GEP3), and (v) (GEP5), where the $R_a$, $T_{max}$, $T_{min}$ and $R_a$, $T_{max}$, $T_{min}$, DOY, ICSKY are introduced as input parameters respectively. Both input combinations present similar performance (validation period): 3.33 MJ m$^{-2}$ d$^{-1}$ (RMSE) and 15.58 ($r^2_{RMSE}$) in the case of GEP3; 3.31 MJ m$^{-2}$ d$^{-1}$ (RMSE) and 16.01 ($r^2_{RMSE}$) in the case of GEP5. So Table 2 clearly shows that the GEP model is not much sensitive to DOY and ICSKY input variables. However, the input combination (iv) ($R_a$, $T_{max}$, $T_{min}$, DOY) is clearly the optimal combination for ANFIS and ANN models (ANFIS4 and ANN4) with relatively low error values: 3.14 MJ m$^{-2}$ d$^{-1}$ (RMSE), and 20.40 ($r^2_{RMSE}$) in the case of ANFIS4; 2.93 MJ m$^{-2}$ d$^{-1}$ (RMSE) and 25.60 ($r^2_{RMSE}$) in the case of ANN4.

ANN4 presents the best performance among the studied models. ANN4 is a multilayer perceptron (MLP) with ten neurons in the hidden layer. In Fig. 2 there is a description of this model. The MLP formulation has the ability to explore a wide range of multivariable regressions in a single step, and the non-linear property of its activation function is valuable for reproducing the behavior of a natural phenomenon like the relation between global solar radiation and other meteorological variables. As it was mentioned in the introduction section many authors have used the ANN based global solar radiation estimation approach. And the performance obtained by ANN4 in terms of RMSE (2.93 MJ m$^{-2}$ d$^{-1}$) is concordant with the obtained by other authors. Bocco et al. [35] obtained RMSE values from 3.15 to 3.88 MJ m$^{-2}$ d$^{-1}$ in Argentina for different ANN models. Liu and Scott [36] obtained RMSE values varying between 2.01 and 5.44 MJ m$^{-2}$ d$^{-1}$ in Australia. Fortin et al. [10] in their study in Canada, obtained RMSE values about 3.8 MJ m$^{-2}$ d$^{-1}$ based on a four input ANN.

ANFIS4 is an interesting alternative to ANN4 (3.14 MJ m$^{-2}$ d$^{-1}$ of RMSE). There are not many studies of $R_S$ estimation based on ANFIS. Mellit et al. [14] obtained an ANFIS with the ability of modeling $R_S$ based on mean temperature and sunshine hour's inputs. The present study demonstrates the ability of ANFIS models to model global solar radiation based on temperatures and extraterrestrial radiation.

The accuracy of GEP models seems to be slightly lower than the ANFIS and ANN models. However this study reports a new and efficient approach for the formulation $R_S$ using GEP for the first time. The Genetic Programming models (i.e., GEP) are superior to other artificial intelligence models in giving simple explicit equation (Ta-

**Table 3**
Gene Expression Programming models equations.

| GEPs | Equations |
|---|---|
| GEP1 | $R_S = \sqrt[3]{(R_a - 14.1)^2} + [\cos(\sqrt{R_a} + 0.87)] + \sqrt[3]{\left[arctg(\sqrt{R_a}) - 0.023R_a^2\right]}$ |
| GEP2 | $R_S = -\frac{T_{max}}{R_a}(SinR_a + 9.214) + \left[arctg\left(\left(1 - \frac{T_{max}}{R_a}\right)^2 R_a\right)\right] + \sin[\sin(19.31 + \sqrt{R_a})] + T_{max}$ |
| GEP3 | $R_S = 0.003R_a \cdot \cos\sqrt{R_a} + T_{max} - \frac{R_a \cdot T_{max} + 9.916T_{max}}{1.93R_a} + Ln\left[(arctg(0.802 \cdot \sqrt[3]{R_a})^3\right]$ |
| GEP4 | $R_S = [arctg(0.997T_{min})]^3 + arctg[0.18T_{max} - T_{min} + 4.987] + 0.412(R_a - T_{max}) - T_{max} - 7.783$ |
| GEP5 | $R_S = 0.99\left(\cos\sqrt{ICSKY} - T_{min}\right) + \sqrt{arctg(ICSKY) \cdot (T_{max}^2 + 0.952R_a)} + \frac{\cos(-0.22ICSKY) + 5.45}{ICSKY}$ |

*Note*: $R_S$, Solar radiation; $R_a$, extraterrestrial radiation; $T_{mean}$, daily mean temperature; $T_{max}$, daily maximum temperature; $T_{min}$, daily minimum temperature; *DOY*, day of the year; *ICSKY*, corrected clear-sky solar irradiation.

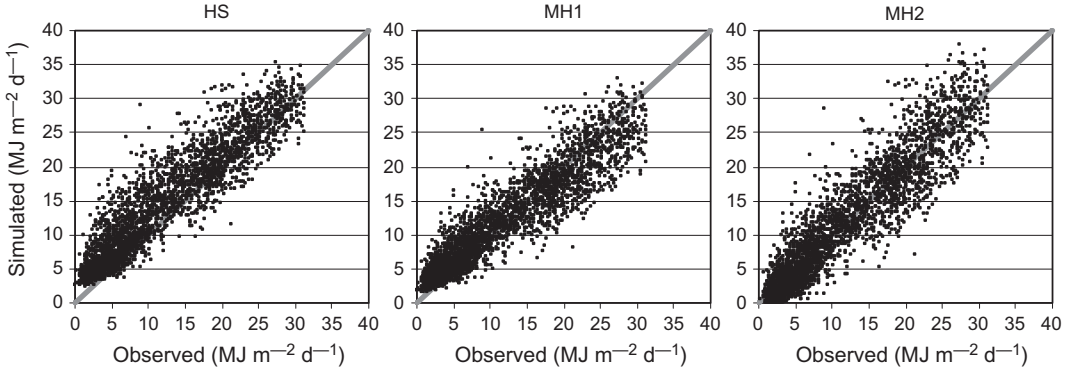


**Fig. 5.** Validation scatter plots of the observed and simulated solar radiation for HS, MH1 and MH2 models.

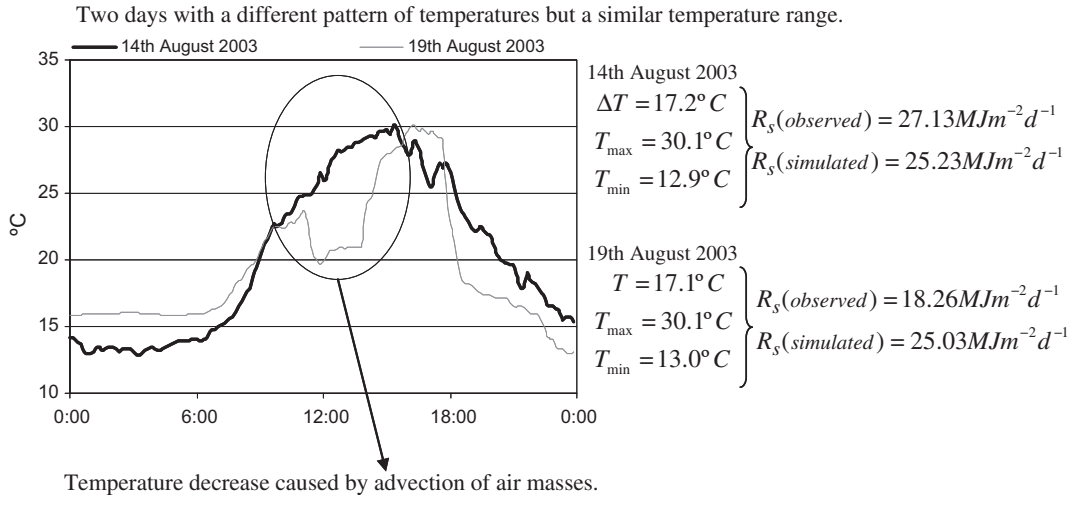


**Fig. 6.** Example of disturbances that affect the relation between solar radiation and temperature.

ble 3) for the phenomenon which shows the relationship between the input and output parameters.

### *4.4. Temperature based global solar radiation estimation procedures performance*

The relationship between global solar radiation and daily temperature range is based on the assumption that global solar radiation is major driver of temperature along the day. On reaching to the earth's surface, net incoming radiation can be partitioned between sensible and latent heat. And sensible heat causes the elevation of daytime temperature along the day, above dose existing under night-time conditions [37]. But there are other factors that determine the temperature throughout the day: wind speed, air vapor content, availability of soil water for evaporation, elevation, precipitation, large-scale advection of air masses, and mesoscale weather phenomena including convective storms. All these factors introduce disturbances that affect the relation between global solar radiation and temperature, and affect the estimation of global solar radiation based on temperature range. So the performance of any procedure of estimation of global solar radiation based on temperature range will depend on the incidence of these disturbance factors. Fig. 6 is an example of this. This figure shows the daily temperature values of 2 days of the same week (Arkaute station). The temperature range of the 2 days is the same, and the HS simulated global solar radiation is almost the same in the two

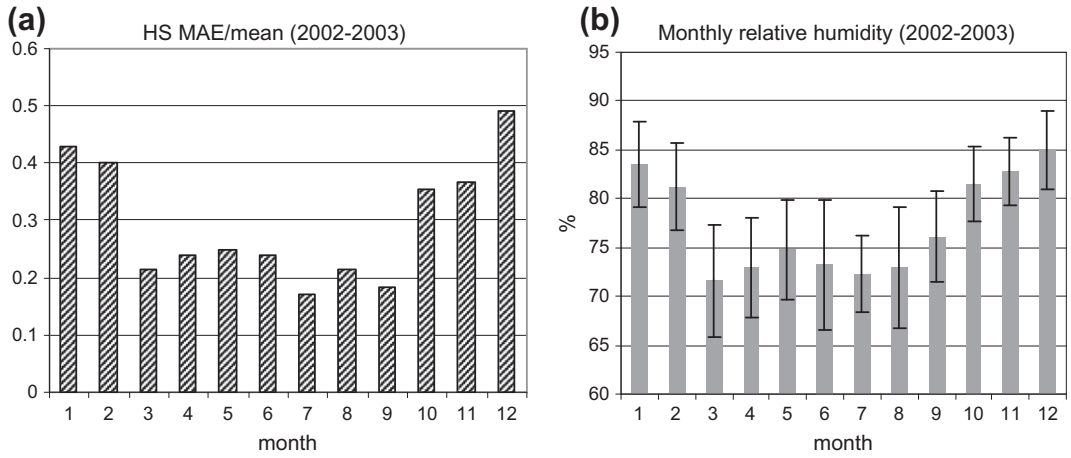


**Fig. 7.** Monthly MAE/mean solar radiation values of HS model (a), and monthly relative humidity values (b) (pool of the four stations, period 2002–2003).

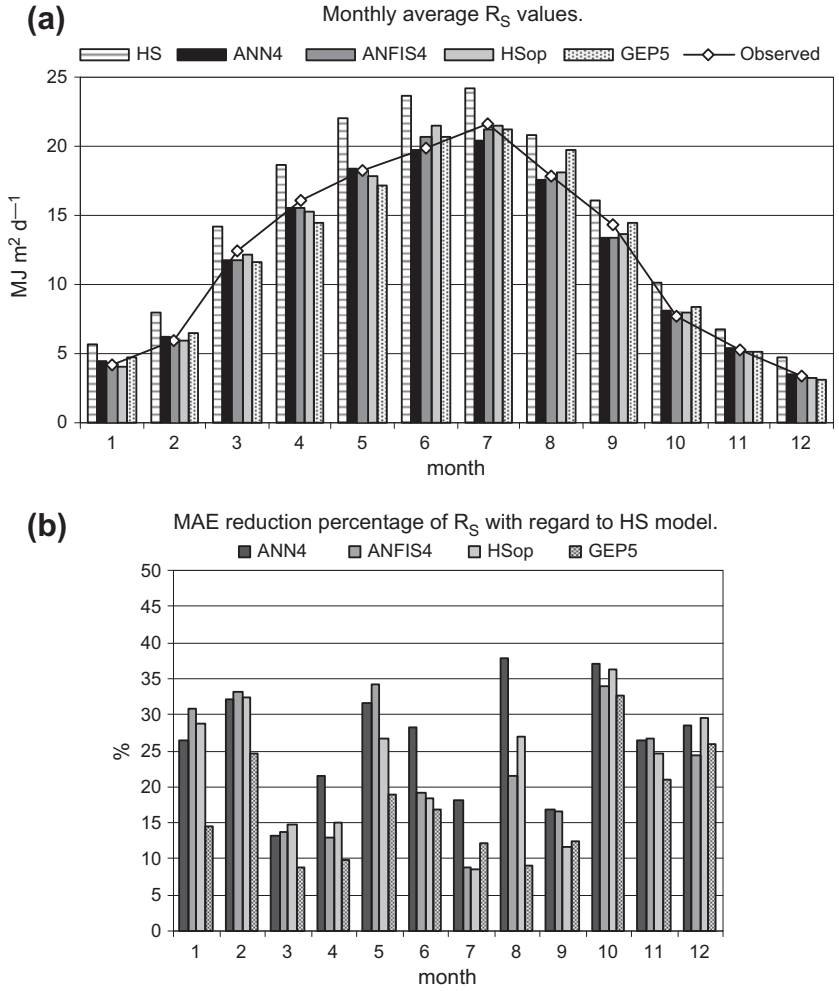


**Fig. 8.** Monthly average $R_S$ values (a) and monthly MAE reduction percentage of estimation of $R_S$ with regard to HS model (b) for the selected models. Evaluation period (2002–2003).

cases. Nevertheless in one of the cases the maximum temperature of the day is reached after a cooling event of some hours caused by a disturbance factor. So despite having the same temperature range, the days shown in Fig. 6 present a different structure of temperatures along the day and different values of observed global solar radiation.

The area of study, as it was mentioned above, is a relative humid location with high $\Delta T$ (daily temperature range) values and low wind speeds. Under these conditions relative humidity of the air seems a good indicator of the performance of a global solar radiation estimation procedure based on temperature range, such as Hargreaves Samani equation (the reference equation of this study).

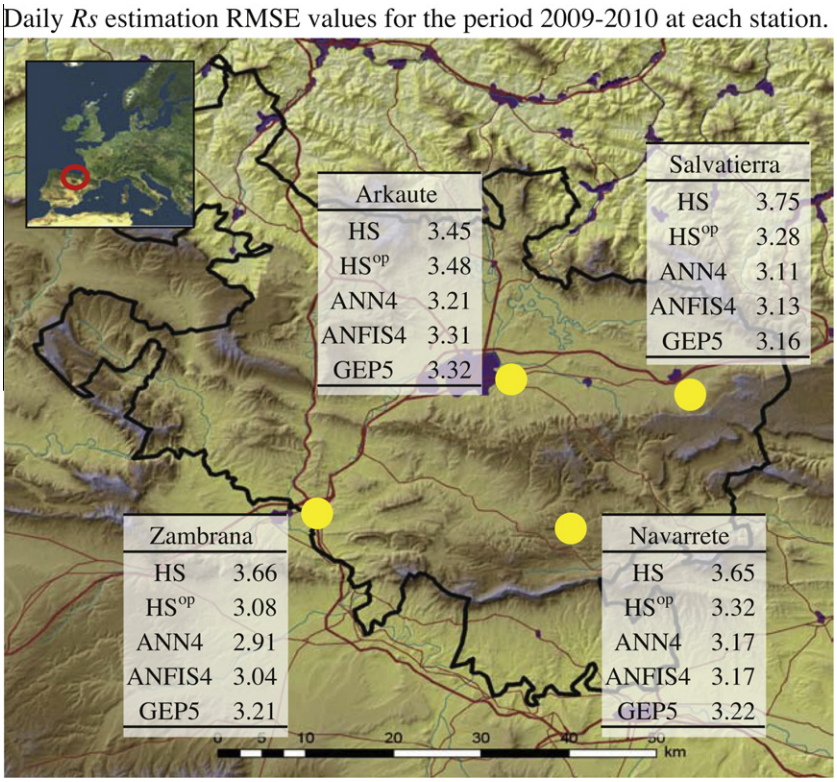


**Fig. 9.** Practical application: daily $R_S$ estimation RMSE values for the period 2009–2010 at each station.

Fig. 7a shows the monthly MAE/mean global solar radiation values of HS model. The use of the MAE/mean as an index of evaluation of the performance of a model is recommended by Kolassa and Schutz [38]. According to Fig. 7a the performance of HS equation, the reference temperature based global solar radiation estimation model, is better from March to September than for the rest of the year. This better performance seems related with lower values of relative humidity of the air along this period (Fig. 7b). The monthly MAE/mean values of estimation of global solar radiation for HS model are closely related with the monthly mean relative humidity values presented in Fig. 7b (validation period 2002–2003). Relative humidity is related with air vapor content and availability of soil water for evaporation which are factors that determine the temperature throughout the day. So in situations of high relative humidity values, the amount of net incoming radiation partitioned to latent heat increases, decreasing the amount partitioned to sensible heat. This causes disturbances in the relation between daily global solar radiation and daily temperature, and the performance of temperature based global solar radiation estimation procedures decreases.

Fig. 8 shows the monthly percentages of reduction, with regard to HS equation, of the error of estimation of global solar radiation for the best models of each one of the procedures of estimation evaluated in this study: non-linear optimization ($HS^{opt}$), ANNs (ANN4), ANFIS (ANFIS4), and GEPs (GEP5). All these models significantly improved the performance of estimation of global solar radiation with regard to HS model. Attending to this figure the improvements in the performance associated with these procedures are obtained, moreover, in the period from October to February. As it was shown in Fig. 7b this period is characterized by higher values of relative humidity which decreases the performance of any temperature based global solar radiation estimation approach. These global solar radiation estimation procedures have the ability to take into account, better than the HS model, the increments of incoming global solar radiation partitioned to latent heat. In the case of the period from March to September, the disturbances of the daily temperature values related with advection of air masses or mesoscale weather phenomena are more significant. During this period the non-linear procedures evaluated, despite reducing the errors of estimation of global solar radiation with regard to HS equation, show lower percentages of reduction of these errors. Nevertheless ANN4, even during this period, shows high values of reduction of the errors of estimation of global solar radiation with regard to HS model. These error reduction values are significantly higher than the values of the rest of the procedures, in June, July and August. During these months the incidence of mesoscale weather disturbance factors, such as convective storms, is more common. So ANNs show a higher ability to take into account these kinds of disturbances than the rest of the models.

As it was mentioned above, this study reports a new approach for the formulation of global solar radiation using GEP for the first time. According to Fig. 8 GEP5 clearly improves the performance of HS model. Despite the performance of the other three non-linear procedures is slightly better than the performance of GEP5, GEP modeling has revealed as a new interesting approach for the estimation of global solar radiation based on temperatures. It is worth noting that empirical global solar radiation formulations developed to date are mostly based on predefined functions. However in the GEP approach there is no predefined function to be considered (GEP randomly creates formed functions and selects the one that best fits the experimental results). Moreover, there is no restriction in the complexity and structure of the randomly formed functions. And compared with ANFIS and ANNs, GEPs are much more explicit and simple to apply. GEP models have less parameter than those of the ANFIS and ANN models. Optimal GEP model (GEP5) has only five parameters for estimating global solar radiation. Optimal ANN model (ANN4) has 4, 10 and 1 nodes for the input, hidden and output nodes, respectively (see Fig. 2). It has 61

parameters ($4 \times 10 + 10 + 11$(for biases) = 61). Optimal ANFIS model (ANFIS4) has much more parameters than the optimal ANN4 model. ANFIS4 model has 13 triangular membership functions (MFs) for the four inputs and each MF has three parameters. The total number of premise parameters is 39 and the number of rules is $3^3 \times 4 = 108$. The number of consequent parameters is 108.

As a practical application of this study, 2009–2010 $R_S$ daily values were estimated based on the $R_S$ estimation regional models developed in the present study (Fig. 9). RMSE values shown in Fig. 9 are in accordance with the performance of the models for the implementation and evaluation periods.

## 5. Conclusions

This study compared the performance of classical empirical daily incoming surface global solar radiation equations (Hargreaves and Samani [15] equation, and two versions of Mahmood and Hubbard [16] equation), and different artificial intelligence approaches (ANNs, ANFIS, and GEPs (a new and efficient approach for the formulation of global solar radiation)) using data from four meteorological stations representative of the climate of a Northern region of Spain. The empirical equations were compared with the artificial intelligence models using the original equations, and using an optimized version of these equations. Among the empirical equations the original Mahmood and Hubbard equation (MH1) (3.21 MJ $m^{-2}$ $d^{-1}$ of RMSE) and the optimized Hargreaves Samani equation ($HS^{op}$) (3.14 MJ $m^2$ $d^{-1}$ of RMSE) clearly improved the performance of the original HS equation, the reference model in this study. The ANN4 (a four-input multilayer perceptron with ten neurons in the hidden layer) presented the best performance among the studied models with a RMSE of 2.93 MJ $m^{-2}$ $d^{-1}$. ANFIS4 (a four-input ANFIS model) revealed as an interesting alternative to ANN4 (3.14 MJ $m^{-2}$ $d^{-1}$ of RMSE). Very limited number of studies has been done on estimation of global solar radiation based on ANFIS. And the present one demonstrated the ability of ANFIS to model global solar radiation based on temperatures and extraterrestrial radiation. By the way this study demonstrated, for the first time, the ability of GEP models to model global solar radiation based on daily atmospheric variables. Despite the accuracy of GEP models was slightly lower than the ANFIS and ANN models (3.31 MJ $m^{-2}$ $d^{-1}$ of RMSE for GEP5), the Genetic Programming models (i.e., GEP) are superior to other artificial intelligence models in giving a simple explicit equation for the phenomenon which shows the relationship between the input and output parameters. The GEP models are explicit and simple such that they can be used, by anyone not necessarily being familiar with GEP, in a spreadsheet on a very simple PC, even on a hand-held calculator.

As a general conclusion, this study provided a battery of alternatives for the estimation of global solar radiation values, including a novel approach such as Gene Expression Programming.

## References


[1] Droogers P, Allen RG. Estimating reference evapotranspiration under inaccurate data conditions. Irrig Drain Syst 2002;16(1):33–45.
[2] Bechini L, Ducco G, Donatelli M, Stein A. Modelling interpolation and stochastic simulation in space and time of global solar radiation. Agric Ecol Environ 2000;81:29–42.
[3] Maghrabi AH. Parameterization of a simple model to estimate monthly global solar radiation based on meteorological variables, and evaluation of existing solar radiation models for Tabouk, Saudi Arabia. Energy Convers Manage 2009;50:2754–60.
[4] Robaa SM. Validation of the existing models for estimating global solar radiation over Egypt. Energy Convers Manage 2009;50:184–93.
[5] Li MF, Liu HB, Guo PT, Wu W. Estimation of daily solar radiation from routinely observed meteorological data in Chongqing, China. Energy Convers Manage 2010;51:2575–9.
[6] Liu J, Liu Jingmiao, Linderholm HW, Chen D, Yu Q, Wu D, et al. Observation and calculation of the solar radiation on the Tibetan Plateau. Energy Convers Manage 2012;57:23–32.
[7] Shiri J, Kisi O, Landeras G, López JJ, Nazemi AH, Stuyt LCPM. Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). J Hydrol 2012;414–415:302–16.
[8] Elizondo D, Hoogenboom G, Mcclendon RW. Development of a neural network model to predict daily solar radiation. Agric Forest Meteorol 1994;71:115–32.
[9] Sozen A, Arcaklioglu E, Ozalp M. Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. Energy Convers Manage 2004;45:3033–52.
[10] Fortin JG, Anctil F, Parent LE, Bolinder MA. Comparison of empirical daily surface incoming solar radiation models. Agric Forest Meteorol 2008;148:1332–40.
[11] Lam JC, Wan KKW, Yang L. Solar radiation modelling using ANNs for different climates in China. Energy Convers Manage 2008;49:1080–90.
[12] Benghanem M, Mellit A, Alamri SN. ANN-based modelling and estimation of daily global solar radiation data: a case study. Energy Convers Manage 2009;50:1644–55.
[13] Marti P, Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage 2011;52:990–1003.
[14] Mellit A, Hadj Arab A, Khorissi N, Salhi H. An ANFIS-based forecasting for solar radiation data from sunshine duration and ambient temperature. In: IEEE power engineering society general meeting, 24–28, June, Florida, USA; 2007. p 1–6.
[15] Hargreaves GH, Samani ZA. Estimating potential evapotranspiration. J Irrig Drain Eng 1982;108(3):225–30.
[16] Mahmood R, Hubbard KG. Effect of time of temperature observation and estimation of daily solar radiation for the Northern Great Plains, USA. Agron J 2002;94(4):723–33.
[17] Cengiz SH, Gregory JM, Seabaugh JL. Solar radiation prediction from other climatic variables. Trans ASAE 1981;24:1269–72.
[18] Fausset LV. Fundamentals of neural networks: architectures, algorithms and applications. Upper Saddle River (NJ): Prentice Hall; 1994.
[19] Bishop CM. Neural networks for pattern recognition. Oxford University Press; 1995.
[20] Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River (New Jersey): Prentic-Hall; 1999. ed..
[21] Jang JSR. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst Manage Cybern 1993;23(3):665–85.
[22] Jang JSR, Sun CT, Mizutani E. Neurofuzzy and soft computing: a computational approach to learning and machine intelligence. New Jersey: Prentice-Hall; 1997.
[23] Mamdani EH, Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man Mach Stud 1975;7(1):1–13.
[24] Takagi T, Sugeno M. Fuzzy identification of systems and its application to modeling and control. IEEE Trans Syst Man Cybern 1985;15(1):116–32.
[25] Koza JR. Genetic programming: on the programming of computers by means of natural selection. Cambridge: The MIT Press, Bradford Book; 1992.
[26] Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Reading: Addison-Wesley; 1989.
[27] Banzhaf W, Nordin P, Keller PE, Francone FD. Genetic programming. San Francisco (CA): Morgan Kaufmann; 1998.
[28] Ferreira C. Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst 2001;13(2):87–129.
[29] Softonic, SPSS release 17.0, Computer program. Softonic, Barcelona, Spain; 2008. <http://www.softonic.com> [16.11.11].
[30] Inc. StatSoft, Statistica Neural Networks release 4.1, Computer program. StatSoft Inc., Tulsa, Oklahoma, USA; 1999. <http://www.statsoft.com> [16.11.11].
[31] Inc. Gepsoft, GeneXproTools release 4.0, Computer program. Gepsoft, Inc., Bristol, United Kingdom; 2006. <http://www.gepsoft.com> [16.11.11].
[32] Nash JE, Sutcliffe JE. River flow forecasting through conceptual models. J Hydrol 1970;10:282–90.
[33] Liu XY, Mei XR, Li YZ, Wang QS, Jensen JR, Zhang XQ. Evaluation of temperature-based global solar radiation models in China. Agric Forest Meteorol 2009;149:1433–46.
[34] Russel SO, Campbell PF. Reservoir operating rules with fuzzy programming. J Water Resour Plan Manage 1996;122(3):165–70.
[35] Bocco M, Ovando G, Sayago S. Development and evaluation of neural network models to estimate daily solar radiation at Córdoba. Argent Pesq Agropec Bras 2006;41:179–84.
[36] Liu DL, Scott BJ. Estimation of solar radiation in Australia from rainfall and temperature observations. Agric Forest Meteorol 2001;106:41–59.
[37] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric Forest Meteorol 1984;31:159–66.
[38] Kolassa S, Schutz W. Advantages of the MAD/MEAN ratio over the MAPE. Int J Appl Forecast 2007;6:40–3.