AUTHOR QUERY FORM

ELSEVIER	**Journal: ECM**	**Please e-mail or fax your responses and any corrections to:**
		E-mail: corrections.esch@elsevier.sps.co.in
	Article Number: 4289	**Fax: +31 2048 52799**

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

No queries have arisen during the processing of your article.

Thank you for your assistance.

Energy Conversion and Management xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance

Pau Martí [a,*], María Gasque [b]

[a] Departamento de Ingeniería Rural y Agroalimentaria, Universidad Politécnica de Valencia. Camí de Vera s/n, 46022 Valencia, Spain
[b] Departamento de Física Aplicada, Universidad Politécnica de Valencia. Camí de Vera s/n, 46022 Valencia, Spain

A R T I C L E I N F O

Article history:
Received 23 December 2009
Received in revised form 13 August 2010
Accepted 25 August 2010
Available online xxxx

Keywords:
Artificial neural networks
Solar radiation
Exogenous variables

A B S T R A C T

The development of new and more precise temperature-based models for solar radiation estimation is decisive, given the immediacy and simplicity associated in their input measurements and the ubiquitous problems derived from equipment failures, maintenance and calibration, and physical and biological constraints. Further, the performance quality of empirical equations is to be questioned in a large variety of climatic contexts. As an alternative to traditional techniques, artificial neural networks (ANNs) are highly appropriate for the modelling of non-linear processes. Nevertheless, temperature-based ANN models do not always provide accurate enough solar radiation estimations as their performance depends considerably on the specific temperature/solar radiation relationships of the studied context. This paper describes a new procedure to improve the performance accuracy of temperature-based ANN models for estimation of total solar radiation on a horizontal surface (R_s) taking advantage of ancillary data records from secondary similar stations, which work as exogenous inputs. The influence on the model performance of the number of considered ancillary stations and the corresponding number of training patterns is also analyzed. Finally, these models are compared with those relying exclusively on local temperature recordings. The proposed models provide performances with lower associated errors than those which do not consider exogenous inputs. The ancillary supply is translated into a decrease around 0.1 of RMSE in the local performance. The consideration of non-measured inputs in the simple local temperature-based models, namely extraterrestrial radiation or day of the year, entails a performance accuracy improvement around 0.1 of RMSE.

1. Introduction

The design and development of energy efficient buildings and solar energy conversion (photovoltaic or solar thermal) systems for a particular studied location and application requires accurate estimations of long-term global solar radiation data to simulate the operating conditions of the system [1,2]. Solar radiation also plays an important role in many physical, biological and chemical processes, such as plant photosynthesis, evaporation or crop growth and productivity [3,4]. It is also necessary in biophysical models for risk assessment of forest fires, hydrological simulation models of natural processes [5], environmental and agrometeorological research, or atmospheric physics [6].

Total (global) solar radiation is the sum of the beam and diffuse solar radiation on a surface. The most common solar radiation measurements registered in meteorological stations correspond to total radiation on a horizontal surface, R_s [2], normally given on an hourly or daily basis.

Solar based applications are highly interesting in places where no connection to an electrical supply grid is available, like rural, mountainous or remote areas and natural parks, as well as in many developing countries [7–10]. Unfortunately, despite its significance, global solar radiation measurements are generally not available at the places of interest due to the high-cost installation, maintenance and calibration associated to radiometric stations [7,9,11]. Nevertheless, in some cases, there are meteorological stations without solar radiation sensors, where other variables can be registered [5]. Even in automatic meteorological stations where solar radiation is measured, data records are often missing due to equipment failure, erroneous because of sensor calibration problems, or lie outside the expected range [1,12,13].

Therefore, different empirical and numerical models for global terrestrial solar radiation estimation, based on different meteorological input combinations, have been proposed for those cases where radiation data are not available [2,6,7,14]. The different solar radiation models differ in sophistication from simple empirical formulations based on common climate data to more complex

* Corresponding author. Tel.: +34 963877521; fax: +34 063877549.
E-mail addresses: paumarpe@doctor.upv.es (P. Martí), mgasque@fis.upv.es (M. Gasque).

0196-8904/$ - see front matter © 2010 Published by Elsevier Ltd.
doi:10.1016/j.enconman.2010.08.027

Nomenclature

b_k	bias	s	number of repetitions
CI	continentality index	T_{max}	daily maximum air temperature
CI^C	Conrad continentality index	T_{mean}	daily mean air temperature
CI^{CU}	Currey continentality index	T_{min}	daily minimum air temperature
CI^G	Gorezynski continentality index	u_2	wind speed at 2 m height
CI^S	Supan continentality index	U_x	maximum value assigned in the scaled sample
e_k	vector of network errors	u_x	minimum value assigned in the scaled sample
ET_o	reference evapotranspiration	v_k	summing junction
$\boldsymbol{I}$	unit matrix	w_{kj}	synaptic weight of neuron k
$\boldsymbol{J}$	Jacobian matrix	x	original variable
l	number of layers	x_k	input signal
M_i	maximum monthly average temperature	x_s	scaled variable
m_i	minimum monthly average temperature	y_e	expected vector
M_x	maximum value of the original sample	y_k	output variable
m_x	minimum value of the original sample	y_m	predicted vector
MBE	mean bias error	Greek symbolsΔT	
MSE	mean squared error		daily temperature range
n	number of hidden neurons	Φ	latitude
r^2	determination coefficient	μ	constant that governs the step size
R_a	extraterrestrial radiation	φ	hyperbolic tangent sigmoid function
RH	air relative humidity	σ_e	expected standard deviation
RMSE	root mean squared error	σ_m	predicted standard deviation
R_s	solar radiation		

numerical models which usually involve high computational costs and also require numerous input parameters.

Another alternative is the application of mathematical models like artificial neural networks (ANNs). ANNs are simplified models of the central nervous system which may be used as effective tools to model non-linear problems. They can be defined as massively parallel distributed processors consisting of simple processing units, which have a natural propensity for storing experimental knowledge and making it available for use [15]. An ANN is configured for a specific application through a learning process. Learning in biological systems as well as in ANNs involves adjustments to the synaptic connections that exist between the neurons. During the last decades, it has taken place an important increase in their application in different scientific areas due to the development of computer technologies.

Among the most common ANN applications are: constraint satisfaction, control, data compression, diagnostics, forecasting, general mapping, multisensory data fusion, optimization, pattern recognition and risk assessment [16]. ANNs can detect more complex properties of the studied data than traditional statistical techniques because of their non-linear structure [17]. Further, they do not require detailed information regarding the physical processes of the system.

ANNs have been successfully applied by many researchers for solar radiation estimation considering different ANN types and input combinations in different parts of the world [4–8,10,14,18–25], including Spain [3,9,26,27].

Nevertheless, only a small part of the aforementioned papers consider a low number of inputs. And among these, only few of them do not consider sunshine duration as input data. Kalogirou et al. [22] proposed a neural network for R_s estimation demanding only measured air temperature and relative humidity records. Rehman and Mohandes [8] analyzed the performance of three ANNs for R_s estimation considering maximum temperature, mean temperature and mean temperature/relative humidity, respectively, as measured inputs. Finally, Benghanem et al. [25] tackled the ANN performance reached with the consideration of air temperature and relative humidity as measured inputs, individually and together, and stated the performance improvement derived from adding in the mentioned ANNs measured sunshine duration as input, too.

Among the simplest methods for estimating historical solar radiation data, Hargreaves and Samani [28], Bristow and Campbell [29], and Allen [30] suggested that solar radiation could be estimated as a function of maximum and minimum temperatures and extraterrestrial radiation (R_a). These empirical methods, modified by other authors [11], consider, implicitly, the particular location of the area and the period of study, as they account for latitude, day of the year, sunset hour angle, or relative distance earth–sun by including R_a inputs.

The development and improvement of temperature-based models can play a decisive role in solar radiation estimation, given the immediacy and simplicity associated in their input measurements and the aforementioned ubiquitous problems derived from equipment failures, maintenance and calibration, and physical and biological constraints. Nevertheless, as could be foreshadowed, temperature-based R_s models present a serious drawback: their accuracy depends considerably on the temperature range (ΔT) of the application area and on the specific local temperature/solar radiation relationships. Larger ΔT generally results in better predictive accuracy [11]. Bearing this in mind, the current study presents a new procedure to improve the performance accuracy of temperature-based ANN models for R_s estimation taking advantage of ancillary data records from secondary similar stations, which work as exogenous inputs. This methodology has been successfully applied in water resources for improving the performance of temperature-based ANNs for reference evapotranspiration (ET_o) estimation [31]. So, first, the most suitable ancillary stations are selected through a continental characterization of the study area. Next, different input combinations are defined, trained and tested. The influence on the model performance of the number of considered ancillary stations and the corresponding number of training patterns is also analyzed. Finally, these ANNs are compared with those models based exclusively on local temperature records.

Please cite this article in press as: Martí P, Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage (2010), doi:10.1016/j.enconman.2010.08.027

Fig. 1. Situation of the studied stations.

2. Materials and methods

2.1. Climatic data management

The historical series of the climatic variables for this study were obtained from 30 weather stations of the Irrigation Technology Service belonging to the Valencian Institute for Agricultural Research (IVIA), Fig. 1. The daily values of maximum, minimum and average temperature, average and maximum wind speed, relative air humidity, solar radiation and sunshine duration were collected by these automatic meteorological stations between January 2000 and December 2007. These years correspond to a climatologically normal period, without sharp or noticeable changes during all of them. Table 1 sums up the geographical information of the studied stations. A climatic characterization of the considered stations is given in Table 2 through the mean and standard deviation of daily average temperature (T_{mean}), daily thermal oscillation (ΔT), daily wind speed at 2 m height (u_2), daily relative humidity (RH), daily solar radiation (R_s), and daily evapotranspiration (ET_o) for the period 2000–2007.

All source data were scaled in the interval [−0.9; 0.9], avoiding the possibility of imposing higher-order precedence by magnitude. So, a higher numerical efficiency is achieved in the application of the training algorithm. This interval was established to avoid the saturation of the neuron output range and the subsequent limitation of the extrapolation ability which involve the intervals [−1; 1] and [0; 1] for tansig and logsig activation functions, respectively. With these latter intervals, the neural network cannot pro-

Table 1
Geographic characterization of the studied locations.

Station name	Code	Latitude (° ′ ″)	Longitude (° ′ ″)	Altitude (m)
Pilar de la Horadada	1	37 52 12N	00 48 37W	77
Altea	2	38 36 20N	00 04 39W	210
Vila Joiosa	3	38 31 46N	00 15 19W	138
Tavernes de Valldigna	4	39 05 47N	00 14 12W	15
Sagunt	5	39 38 57N	00 17 33W	33
Benavites	6	39 44 00N	00 12 54W	8
Ondara	7	38 49 11N	00 00 27E	49
Denia-Gata	8	38 47 38N	00 05 01E	102
Vall d'Uixó	9	39 47 51N	00 13 38W	100
Vila Real	10	39 56 00N	00 06 00W	42
Almoradí	11	38 05 27N	00 46 17W	74
Moncada	12	39 37 11N	00 20 56W	35
Elx	13	38 16 00N	00 42 00W	86
Sant Rafel del Riu	14	40 35 44N	00 22 13E	205
Catral	15	38 09 16N	00 48 15W	27
Agost	16	38 25 40N	00 38 36W	345
Vilanova de Castelló	17	39 04 00N	00 31 22W	58
Carcaixent	18	39 07 00N	00 30 17W	35
Monforte del Cid	19	38 23 59N	00 43 44W	244
Carlet	20	39 30 00N	00 26 00W	35
Castalla	21	38 36 19N	00 40 22W	708
Orihuela	22	38 10 58N	00 57 13W	99
Turís	23	39 24 02N	00 41 01W	299
Pedralba	24	39 34 04N	00 42 59W	200
Lliria	25	39 41 31N	00 37 31W	250
Cheste	26	39 31 18N	00 44 30W	323
El Pinós	27	38 25 43N	01 03 34W	606
Camp de Mirra	28	38 40 49N	00 46 18W	627
Villena	29	38 35 48N	00 52 24W	495
Campo Arcís	30	39 26 04N	01 09 39W	584

duce output values beyond the maximum considered in the data set. After the simulation, outputs were returned to original values. For this purpose,

$$x_s = \frac{(U_x - u_x)\cdot x + (M_x \cdot u_x - m_x \cdot U_x)}{M_x - m_x}$$

where x_s is the scaled variable; x is the original variable; M_x is the maximum value of the original sample; m_x is the minimum value of the original sample; U_x is the maximum value assigned in the scaled sample; u_x is the minimum value assigned in the scaled sample.

In each station, the daily data series from 2006 to 2007 were used for cross-validating and testing, respectively, the rest were used for training. Despite the random and fluctuating character of climatic variables, the series assignment for training, cross-validating and testing was established chronologically, which is a common practice in the ANN community.

2.2. Continental characterization of studied locations

The proposed models consider two types of variables: local variables, corresponding to the training station, and exogenous variables, corresponding to ancillary stations, climatologically similar to the training station. The criterion used to identify the most appropriate ancillary data-supplier stations was based on a continental characterization of the study region. Therefore, different continentality indexes were calculated for the studied stations. More specifically, the selected indexes were Gorezynski, Conrad, Supan and Currey indexes. These indicators were selected for their simplicity, as they only demand temperature and latitude records. Thus, these were calculated as follows [32]:

$$\mathrm{CI}^{G} = 1.7\frac{M_i - m_i}{\sin(\Phi)} - 20.4$$

$$\mathrm{CI}^{C} = 1.7\frac{M_i - m_i}{\sin(\Phi + 10)} - 14$$

$$\mathrm{CI}^{S} = M_i - m_i$$

$$\mathrm{CI}^{CU} = \frac{M_i - m_i}{1 + \frac{\Phi}{3}}$$

Table 2
Climatic characterization of stations considered. Daily mean values corresponding to the period 2000–2007.

Station code	T_{mean} (°C)		ΔT (°C)		u_2 (m/s)		RH (%)		R_s (W/m^2)		ET_o (mm)	
	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
1	18.09	5.51	9.19	2.84	1.78	0.95	65.58	12.65	201.37	89.62	3.48	1.69
2	18.00	5.64	8.98	2.18	1.17	0.30	61.68	11.91	194.97	92.32	3.19	1.73
3	18.12	5.62	8.75	2.03	1.32	0.39	60.18	12.45	188.25	84.82	3.23	1.60
4	17.73	5.76	9.38	3.24	1.69	0.78	69.31	14.00	184.64	91.50	3.21	1.71
5	17.44	5.92	9.52	3.21	1.36	0.52	62.27	13.41	190.34	90.72	3.14	1.66
6	16.61	5.75	11.58	3.41	1.08	0.45	70.36	11.95	182.79	87.93	2.85	1.50
7	17.49	6.15	11.81	3.89	1.11	0.51	66.73	13.37	179.43	89.01	2.98	1.70
8	16.98	6.09	12.11	3.67	0.86	0.33	69.73	12.57	183.99	90.41	2.80	1.63
9	17.11	5.88	10.12	2.59	1.36	0.32	63.04	13.21	184.81	90.97	3.14	1.62
10	16.55	6.02	10.73	2.62	1.18	0.40	66.17	12.81	178.06	90.38	2.96	1.66
11	18.04	5.72	9.50	2.62	1.42	0.55	65.88	12.46	195.88	87.51	3.33	1.67
12	17.07	6.09	12.15	3.28	1.12	0.65	69.54	12.64	182.81	88.26	3.01	1.68
13	17.01	5.89	10.80	3.01	1.12	0.48	63.38	12.14	190.79	80.36	3.08	1.64
14	15.61	6.20	9.60	2.75	1.63	0.91	65.47	14.75	182.98	93.84	3.05	1.76
15	17.79	6.25	13.55	3.73	1.18	0.65	67.00	11.70	193.47	88.15	3.25	1.74
16	16.28	6.07	10.72	2.96	1.83	0.80	60.47	13.85	195.91	90.07	3.48	1.79
17	17.25	6.69	13.76	4.47	0.90	0.51	68.46	12.35	186.01	95.80	3.01	1.86
18	16.68	6.58	13.58	4.11	0.90	0.41	71.31	12.61	181.52	88.72	2.90	1.79
19	16.66	6.15	11.93	3.40	1.69	0.80	62.37	13.45	185.58	87.08	3.41	1.74
20	16.83	6.34	12.37	4.12	1.34	0.77	69.44	13.29	181.77	88.82	3.10	1.72
21	14.39	6.50	11.19	3.70	2.14	1.05	62.62	14.73	211.96	99.74	3.55	2.05
22	17.91	6.10	11.92	3.31	1.50	0.55	64.10	13.45	204.20	90.44	3.56	1.87
23	16.14	6.09	12.57	4.16	1.50	0.89	65.83	13.34	192.90	95.12	3.23	1.72
24	16.82	6.17	10.92	3.38	1.38	0.77	60.41	14.21	188.79	93.88	3.27	1.79
25	16.12	6.34	13.11	3.78	1.04	0.52	65.00	13.34	190.43	95.15	3.00	1.75
26	16.17	6.15	13.26	4.31	1.09	0.72	63.07	14.19	185.55	91.81	3.00	1.66
27	15.19	6.58	11.21	3.48	2.29	1.02	61.44	14.73	205.84	94.97	3.69	1.96
28	14.62	6.95	12.69	4.16	1.98	0.88	64.49	14.73	194.51	104.21	3.42	2.09
29	14.73	6.89	14.01	4.70	1.92	0.93	65.81	13.13	198.45	92.34	3.46	1.98
30	13.92	7.19	14.51	5.02	1.76	0.84	63.64	14.37	188.31	94.22	3.36	2.02

Please cite this article in press as: Martí P, Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage (2010), doi:10.1016/j.enconman.2010.08.027

Table 3
Model alternatives and corresponding considered inputs.

Model name	Considered inputs	
	Training station	Ancillary station
a_1	T_{max},T_{min}	–
a_2	T_{max}, T_{min}	R_s
b_1	T_{max}, T_{min} (J)	–
b_2	T_{max}, T_{min} (J)	R_s
c_1	T_{max}, T_{min} R_a	–
c_2	T_{max}, T_{min} R_a	R_s

where CI^G is the Gorezynski continentality index (–); CI^C is the Conrad continentality index (–); CI^S is the Supan continentality index (–); CI^{CU} is the Currey continentality index (–); M_i is the maximum monthly average temperature (°C); m_i is the minimum monthly average temperature (°C); Φ is the latitude (degrees).

The values of the aforementioned continentality indexes for each considered station can be found in a recent study in the field of water resources [31]. The four indexes show a very similar trend in the studied region, although they present different ranges. In consequence, the four indexes will lead to the selection of practically the same ancillary stations, as the CI relative differences between stations are quite similar in the four cases [31]. According to the conclusions of this study, only the CI^G was used to select the ancillary data-supplier stations in the present work.

It is important to take into account that these indexes are referred to annual data sets. Thus, the same station presents different CI values each year. Moreover, these fluctuations can be considerable. This raises a question in the selection of the period to which the CI must be referred to. Analytically, two continentality indexes can be considered for each station: one referred to the test year or one mean CI value of the 8 years considered. The first option accounts for the selection of the most similar ancillary stations in the specific climatic context of the test year. So, the ancillary station selection is especially appropriate in the test stage of the model. On the other hand, similarly, following the second criterion, the station selection is especially appropriate in the training stage of the model.

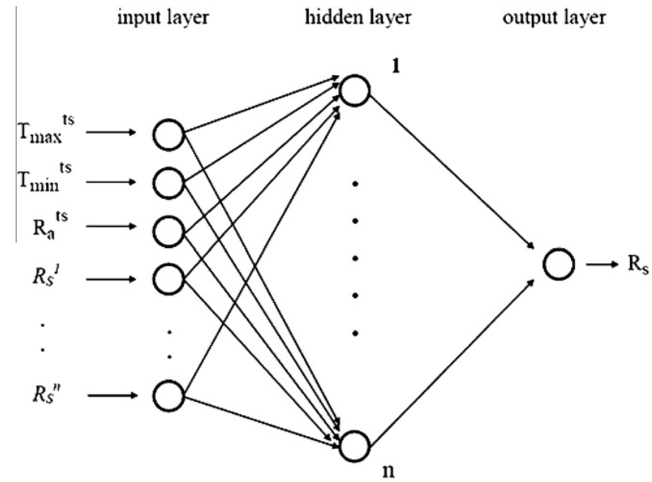

Fig. 3. Architecture scheme of model c_2. Note: exogenous inputs in italics. ts means training station.

The considered climatic series contained data gaps. Thus, if complete monthly data series of any year were missing, the corresponding CI of that year could not be calculated properly attending to their definition. The CI values would not have been reliable, especially if those gaps corresponded to winter or summer, where the extreme temperature records are usually registered. Consequently, stations with monthly gaps could not be considered as ancillary stations for that year, due to the absence of CI values. Moreover, these years were neglected in the calculation of the CI mean value. According to the conclusions of Martí and Gasque [31], only the mean CI was used in the present study. The consid-

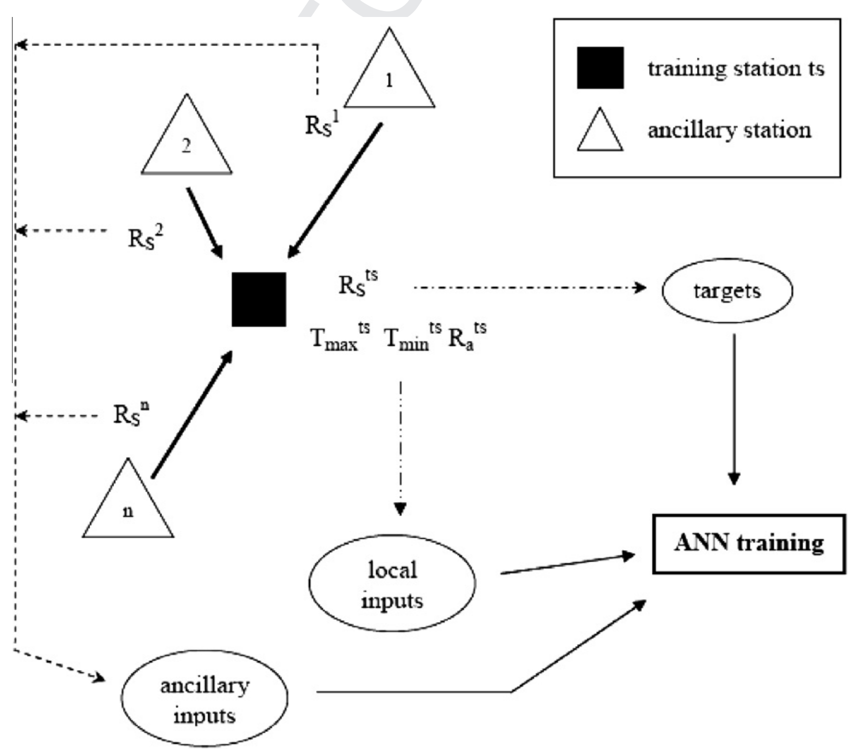

Fig. 2. Diagram of input/output management in model c_2.

Please cite this article in press as: Martí P, Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage (2010), doi:10.1016/j.enconman.2010.08.027

Table 4
Assignment order of ancillary stations according to mean CI.

Training station code	Ancillary station arrangement order														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Ancillary station code														
1	3	2	6	4	11	9	10	5	13	14	12	23	24	19	16
2	6	3	1	4	11	9	10	5	13	14	12	23	24	19	16
3	2	1	6	4	11	9	10	5	13	14	12	23	24	19	16
4	11	9	6	2	3	10	5	1	13	14	12	23	24	19	16
5	10	9	11	13	4	14	12	23	6	2	24	19	16	3	26
6	2	3	4	11	1	9	10	5	13	14	12	23	24	19	16
7	22	25	20	8	15	26	16	19	24	23	12	14	21	13	18
8	26	25	16	19	24	7	23	22	12	14	13	20	15	5	10
9	11	4	10	5	6	2	3	1	13	14	12	23	24	19	16
10	5	9	11	4	13	14	6	12	23	2	24	3	19	16	1
11	9	4	6	10	5	2	3	1	13	14	12	23	24	19	16
12	23	14	24	19	16	13	26	8	5	25	10	7	22	9	11
13	14	12	23	24	19	16	5	10	26	8	9	11	25	4	7
14	12	23	13	24	19	16	26	8	5	10	25	7	22	9	11
15	20	22	7	21	25	8	18	17	26	27	16	19	24	23	12
16	19	24	26	23	12	14	8	13	25	7	22	5	10	20	15
17	18	27	21	29	15	20	22	7	25	8	28	26	16	19	24
18	17	27	21	29	15	20	22	7	25	8	28	26	16	19	24
19	16	24	26	23	12	14	8	13	25	7	22	5	10	20	15
20	15	22	7	25	21	8	26	18	17	16	19	24	23	27	12
21	18	17	27	15	20	22	29	7	25	8	26	16	19	24	23
22	7	25	20	15	8	26	16	19	24	23	12	14	21	13	18
23	12	14	24	19	16	13	26	8	25	5	10	7	22	9	11
24	19	16	23	12	26	14	8	13	25	7	22	5	10	20	15
25	7	22	8	26	20	16	19	24	15	23	12	14	13	21	5
26	16	8	19	24	23	12	14	25	7	13	22	20	15	5	10
27	17	18	21	29	15	20	22	7	28	25	8	26	16	19	24
28	30	29	27	17	18	21	15	20	22	7	25	8	26	16	19
29	27	17	18	21	28	15	20	30	22	7	25	8	26	16	19
30	28	29	27	17	18	21	15	20	22	7	25	8	26	16	19

eration of mean continentality values in the selection of ancillary stations seems to be more appropriate than the consideration of the test year CI, as it might involve a proper selection of the ancillary inputs used in the training stage, which considers a higher amount of data than the test stage. In other words, the selection of ancillary stations will be more realistic and representative of the complete data set. Furthermore, if the CI is referred to the test year, there is higher probability to exclude some stations from the process, because there might exist not enough data for its calculation and, consequently, for the subsequent selection of the corresponding ancillary stations.

2.3. Model alternatives and input management

As pointed out above, the considered models introduce the novelty of taking into account exogenous variables. Accordingly, R_s records can work as targets or as ancillary inputs. In the training station, local R_s values are used as targets whereas R_s values from other stations are used as inputs. Three model types, namely *a*, *b*, and *c*, each one with two alternatives (1 or 2), have been defined attending to the inputs considered. The differences between the three models lie in the consideration or not, respectively, of the day of the year (*J*) values, and the local extraterrestrial radiation

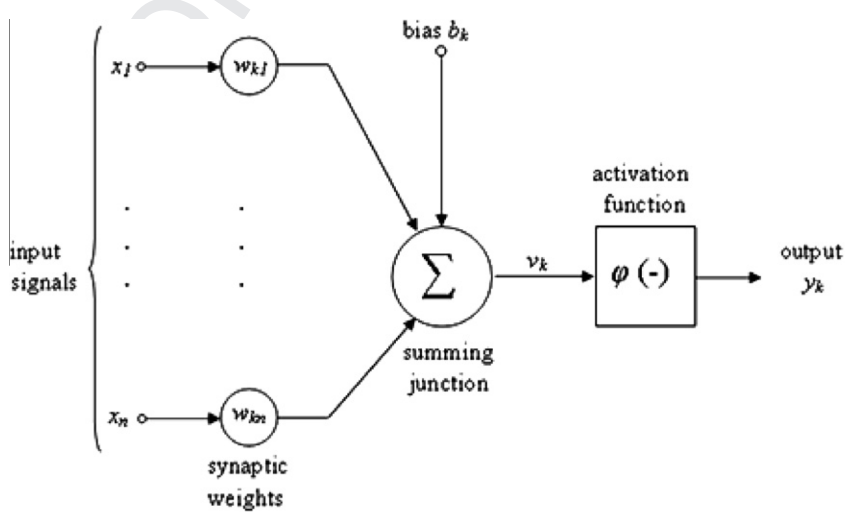

Fig. 4. Configuration of applied neurons.

Table 5
Parameters used in the training process.

Performance function	MSE
Maximum number of epochs to train	100
Performance goal	0
Maximum validation failures	5
Minimum performance gradient	1E-10
Initial , μ	0.001
μ Decrease factor	0.1
μ Increase factor	10
Maximum, μ	1E + 10
Maximum time to train	Infinite

(R_a), which is calculated as a function of the latitude and the day of the year. Model *a* considers only temperature inputs. The difference inside a model pair (1 and 2) lies in the consideration of exogenous R_s records as inputs or not. These model alternatives are summarized in Table 3. *J* is in brackets because this variable is considered as local although it allows no geographical origin assignment.

Each type of model 2 was defined for a number of ancillary stations from 1 up to 15. So, 48 models were performed (15 per model alternative 2 and 3 per model alternative 1) in each station. In model type a_2, the number of ANN inputs ranged between 3 (1 ancillary station) and 17 (15 ancillary stations). In model type b_2 and c_2, the number of inputs ranged between 4 (1 ancillary station) and 18 (15 ancillary stations). Figs. 2 and 3 show the input–output management of model type c_2 and the corresponding translation in a neural network scheme, respectively, where *ts* means training station.

There is a higher probability to incorporate less continentally similar data series to the training set when more ancillary stations are considered. The differences between the training station and the ancillary stations depend on the relationships between the individual CI values of the selected stations, and the CI distribution is not linear [31]. Table 4 sums up the specific ancillary station assignment order that was considered for each training station according to an increasing CI difference.

Every model was tested in the training station (local performance) and in the rest of stations (external performance). Thus, the performance indicators were divided into two groups. First, the local performance was assessed for each model with the local test set. Next, the average external performance in each station was assessed through the mean performance of the remaining 29 models (one per station) in that station [31]. In both cases, these mean results were referred to the number of ancillary stations used.

2.4. ANN configuration and properties

All ANN neurons used were configured, based on the model proposed by Haykin [15]. The neuron of Fig. 4 can be mathematically characterized with the following equations [15]:

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k$$

$$y_k = \varphi(v_k)$$

where x_j is the input signal; w_{kj} is the synaptic weight of neuron k; v_k is the linear combiner or summing junction; b_k is the bias; y_k is the output of the neuron and $\varphi(\cdot)$ is the transfer function. The hyperbolic tangent sigmoid function φ was adopted as activation function. If the output layer of the network has sigmoid neurons, then the output values are limited to a small range. This is why linear output neurons were used, and the network outputs can take on any value.

The ANNs used correspond to multilayer feed-forward networks with back-propagation and supervised training. Thus, they are feed-forward fully-connected hierarchical networks that use differentiable activation functions and supervised training that involves an iterative procedure to minimize the error function (performance function). The errors are used as inputs to feedback connections from which adjustments are made to the synaptic weights layer by layer in a backward direction.

Neural network minimization problems are often very ill-conditioned. This makes the minimization problem harder to solve, and for such problems, the Levenberg–Marquardt algorithm is a good choice. The Levenberg–Marquardt algorithm uses an approximation to the Hessian matrix in the following Newton-like update:

$$x_{k+1} = x_k - [\boldsymbol{J}^T \boldsymbol{J} + \mu \boldsymbol{I}]^{-1} \boldsymbol{J}^T e_k$$

where μ governs the step size and $\boldsymbol{I}$ is the unit matrix; $\boldsymbol{J}$ is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, and e_k is a vector of network errors [33,34]. The selected training parameters are summed up in Table 5. These are standard values for the adopted ANN configuration [35].

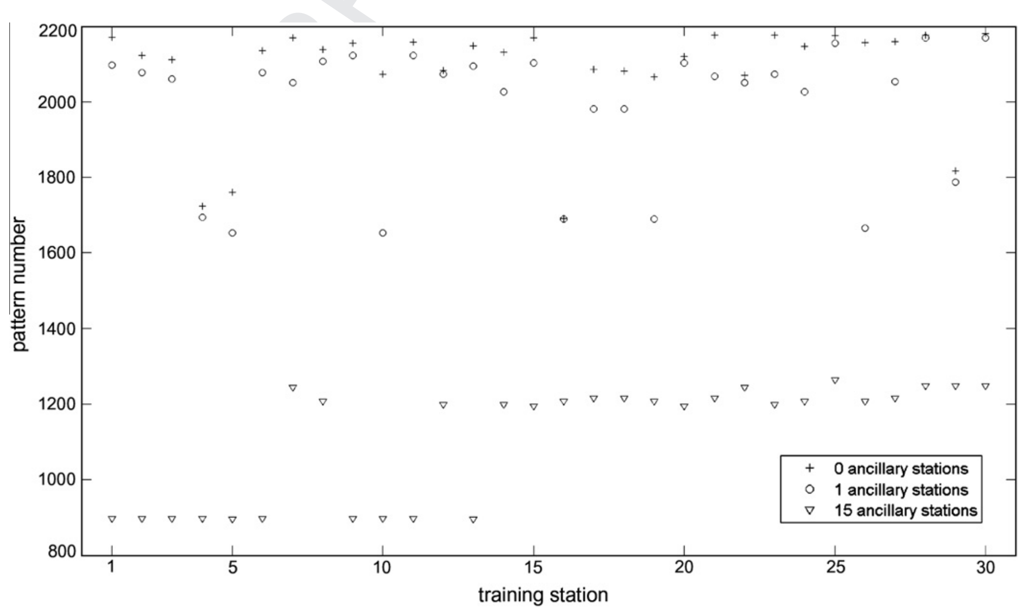

Fig. 5. Reduction of training pattern number per station associated to the homogenization process.

The early stopping procedure was considered to finalize the training. Therefore, training data series were divided into two groups: the first for learning/parameter estimation and the second for cross-validation. The error measured with respect to independent data, the cross-validation set, often shows a decline at first, followed by an increase as the network starts to over-fit [36]. Accordingly, when the chosen error (the MSE) of the cross-validation set was lower than its value in the previous iteration, the training of the network proceeded; otherwise, the training ended. Additional stopping criteria were taken into account, so that training stopped if any of the following conditions were fulfilled:

i. The maximum number of epochs was reached.
ii. The maximum amount of time was exceeded.
iii. Performance was minimized to the goal.
iv. The performance gradient fell below the *minimum performance gradient*.
v. μ exceeded *maximum* μ.

2.5. Model implementation

Instead of following a common methodology among the ANN community, where only several architectures with a fixed number of neurons per layer are defined and tested, a general procedure was developed which allows for the selection of the optimum architecture each time from a set that considers up to l hidden layers with 1 up to n neurons each, where the different hidden layers always present the same number of neurons. Moreover, each architecture is calculated s times and the corresponding ANN parameters are stored, in order to take into account the effects derived from the random assignment of the weights when the training algorithm is initialized. Here, only one hidden layer was considered, due to high number of cases and stations studied. Accordingly, the maximum number of neurons per layer and the number of repetitions were fixed in 20 each. For each architecture the developed program selects the repetition that provides the best performance (in our case the minimum mean squared error) for the cross-validation set of the training station, afterwards selects the architecture with the best cross-validation set performance and, finally, simulates the test data series.

The program allows for the adjustment of the number of stations that provide ancillary data to the training and testing station. Each data point is referred to a day of the year. Thus, the day of the year is used to assemble automatically the input matrices. The time data series differed between stations, due to the presence of data gaps. For every training station, when the number of ancillary stations was fixed, the involved data series had to be homogenized according to the specific days of the year that were simultaneously present in the selected stations. If a data point (a day of the year) of any station considered was missing, that point had to be removed from the other stations involved in the same model training/testing. Fig. 5 presents the pattern number per training station when 0, 1 and 15 ancillary stations are selected according to the mean CI. So, the final pattern reduction can be quantified in each training station. The homogenization process involves an average decrease in the number of training patterns of 1111 data points when 15 ancillary stations are considered.

The program for the ANN application was implemented with Matlab® [35].

2.6. ANN performance indicators

The selected performance function was the measure given by the mean squared error (MSE), defined as

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(y_{\mathrm{m_i}} - y_{\mathrm{e_i}})^2}{n}$$

where y_{m} is the model output and y_{e} the target output. This function was chosen because of its statistical properties and because it is better understood than other measures. It is a non-negative, differentiable function that penalizes large errors more than small ones. Furthermore, the root mean squared error (RMSE, expressed as a fraction), and the mean bias error (MBE, expressed as a fraction) were determined according to

$$\mathrm{RMSE} = \frac{1}{\bar{y}_{\mathrm{e}}}\left(\frac{\sum_{i=1}^{n}(y_{\mathrm{m_i}} - y_{\mathrm{e_i}})^2}{n}\right)^{0.5}$$

$$\mathrm{MBE} = \frac{\sum_{i=1}^{n}(y_{\mathrm{m_i}} - y_{\mathrm{e_i}})}{n\bar{y}_e}$$

Table 6
Average local performance indicators in the 30 training stations.

Number of ancillary stations considered	RMSE (–)	MBE (–)	r^2 (–)	RMSE (–)	MBE (–)	r^2 (–)	RMSE (–)	MBE (–)	r^2 (–)
	Model								
	a_1			b_1			c_1		
	0.3004	0.0182	0.6489	0.1987	0.0164	0.8516	0.1991	0.0089	0.8478
	a_2			b_2			c_2		
1	0.1686	0.0030	0.8860	0.1557	0.0064	0.9070	0.1559	0.0054	0.9060
2	0.1546	0.0038	0.9072	0.1480	0.0040	0.9153	0.1457	0.0027	0.9188
3	0.1389	0.0068	0.9255	0.1361	0.0087	0.9278	0.1353	0.0033	0.9283
4	0.1297	0.0082	0.9361	0.1317	0.0072	0.9330	0.1289	0.0100	0.9371
5	0.1245	0.0037	0.9395	0.1235	0.0048	0.9402	0.1231	0.0062	0.9413
6	0.1165	0.0033	0.9464	0.1184	0.0071	0.9442	0.1158	0.0046	0.9477
7	0.1156	0.0048	0.9467	0.1157	0.0043	0.9467	0.1144	0.0050	0.9476
8	0.1088	0.0086	0.9529	0.1088	0.0106	0.9531	0.1078	0.0098	0.9542
9	0.1034	0.0101	0.9549	0.1059	0.0108	0.9527	0.1033	0.0138	0.9552
10	0.1033	0.0097	0.9539	0.1031	0.0089	0.9551	0.1002	0.0084	0.9563
11	0.1017	0.0099	0.9551	0.1020	0.0085	0.9540	0.1059	0.0084	0.9499
12	0.0992	0.0071	0.9569	0.0998	0.0068	0.9572	0.1009	0.0059	0.9551
13	0.1013	0.0069	0.9546	0.1015	0.0057	0.9557	0.1015	0.0073	0.9557
14	0.1023	0.0075	0.9533	0.1022	0.0085	0.9537	0.1034	0.0058	0.9524
15	0.1033	0.0077	0.9512	0.1036	0.0089	0.9521	0.1019	0.0077	0.9538

Please cite this article in press as: Martí P, Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage (2010), doi:10.1016/j.enconman.2010.08.027

Table 7
Average external performance indicators in the 30 training stations.

Number of ancillary stations considered	RMSE (–)	MBE (–)	r^2 (–)	RMSE (–)	MBE (–)	r^2 (–)	RMSE (–)	MBE (–)	r^2 (–)
	Model								
	a_1			b_1			c_1		
	0.3422	0.0278	0.6029	0.2420	0.0277	0.8229	0.2418	0.0213	0.8192
	a_2			b_2			c_2		
1	0.2088	−0.0006	0.8675	0.2024	0.0063	0.8785	0.2003	0.0045	0.8805
2	0.2016	−0.0029	0.8727	0.2015	0.0011	0.8759	0.1964	−0.0014	0.8812
3	0.1905	−0.0023	0.8803	0.1918	0.0042	0.8802	0.1899	−0.0013	0.8821
4	0.1866	−0.0036	0.8808	0.1914	−0.0040	0.8760	0.1965	0.0087	0.8712
5	0.1859	−0.0030	0.8817	0.1855	−0.0020	0.8813	0.1845	−0.0003	0.8825
6	0.1911	−0.0020	0.8709	0.1864	0.0009	0.8747	0.1829	−0.0005	0.8812
7	0.1831	−0.0017	0.8816	0.1831	−0.0010	0.8798	0.1833	0.0007	0.8795
8	0.1768	0.0005	0.8850	0.1773	0.0051	0.8840	0.1766	0.0054	0.8853
9	0.1733	0.0004	0.8829	0.1740	0.0029	0.8809	0.1742	0.0074	0.8808
10	0.1759	0.0039	0.8782	0.1750	−0.0004	0.8798	0.1714	0.0013	0.8826
11	0.1734	0.0039	0.8810	0.1721	0.0013	0.8815	0.1820	0.0018	0.8688
12	0.1724	0.0017	0.8804	0.1751	0.0024	0.8782	0.1746	0.0036	0.8771
13	0.1755	0.0028	0.8762	0.1736	0.0038	0.8806	0.1762	0.0046	0.8759
14	0.1768	0.0048	0.8724	0.1760	0.0034	0.8744	0.1756	0.0049	0.8729
15	0.1753	0.0047	0.8737	0.1727	0.0051	0.8767	0.1752	0.0021	0.8739

Apart from the mentioned errors, the determination coefficient r^2 was calculated as follows:

$$r^2 = \left(\frac{\operatorname{cov}(y_e, y_m)}{\sigma_e \sigma_m}\right)^2$$

where y_m and y_e are the predicted and the expected outputs, respectively; σ_e, σ_m are the standard deviations corresponding to y_m and y_e; $\bar{y}$ is the average of the corresponding y values.

3. Results and discussion

The performance quality of the proposed models, when they are tested in the training station is gathered in Table 6. Each element of the table corresponds to the mean value of the 30 stations studied. The model a_2, b_2, and c_2 average indicators are arranged according to the number of ancillary stations considered. Comparing the performance of the models without ancillary supply, it can be seen that the consideration of extraterrestrial radiation and day of the year, respectively, allows a marked improvement in the models b_1 and c_1 (RMSE of 0.3004 in model a_1 vs 0.1987 in b_1 and 0.1991 in c_1). Thus, it is possible to improve a temperature-based ANN by considering an extra input which does not demand experimental measurements.

The accuracy of models a_2, b_2 and c_2 depends on the number of ancillary stations considered, presenting a RMSE range between 0.16 and 0.1. In general, the accuracy of these models improves when the number of ancillary stations increases. Nevertheless, the performance quality decreases with more than 12 (models a_2, b_2) and 10 (model c_2) secondary stations. There might be two reasons for this trend. Firstly, the more ancillary stations are considered, the more different might be these stations to the training station from a continental point of view. As highlighted in Section 2, the secondary stations were arranged for a specific training station according to an increasing CI difference. Secondly, due to the homogenization process established to face the data gap problem in the input and output matrix assembly, the number of training patterns is lower the more ancillary stations are considered. So,

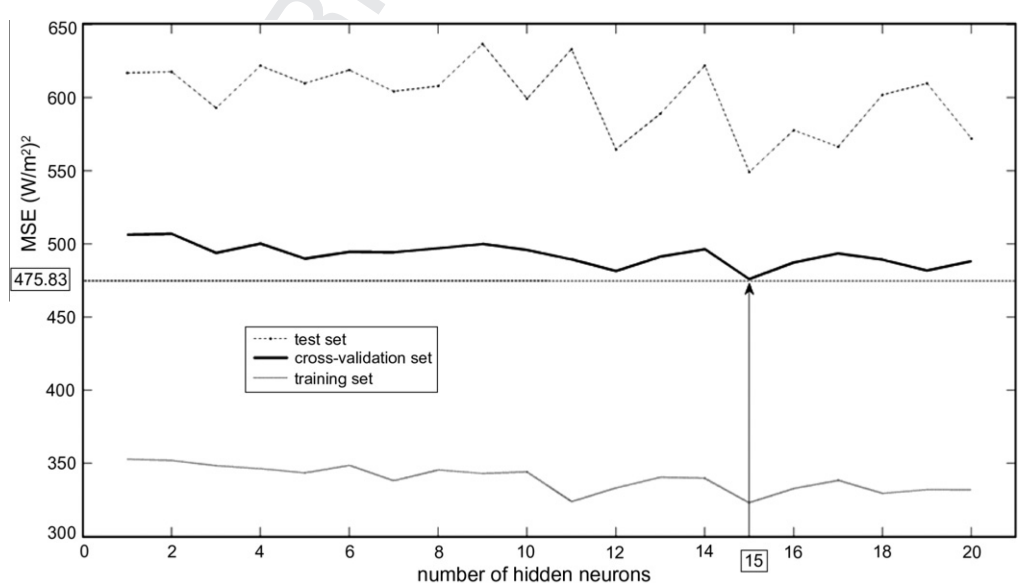

Fig. 6. Optimum architecture selection of model a_2 with 12 ancillary inputs in station 30.

with more than 10–12 ancillary stations, the number of patterns might begin to be insufficient to carry out a proper training.

As observed, models a_2, b_2 and c_2 show a very similar trend in their performance and the indicator differences between them are quite small. This might be due to a higher correlation between local R_s and external R_s than between local R_s and the considered local inputs (T_{max}, T_{min}, R_a and J). Further, the differences between them decrease the more ancillary stations are considered, because the ancillary inputs are the same in the three cases. Hence, according to these results (relative RMSE), the consideration of ancillary R_s data can be translated into an improvement of the model accuracy of 20% when only temperature local records are considered and of 10% when extraterrestrial radiation or day of the year are also considered as inputs. The MBE values show that all these models tend to overestimate R_s. The RMSE reduction achieved with the consideration of the first ancillary station is around 4% in models *b* and *c* and 14% in model *a*. This fact justifies the consideration of a low number of ancillary stations even if only scant secondary stations are available. Similar conclusions can be drawn on the basis of r^2 results, where optimum average values around 0.95 are reached. Due to the aforementioned similarity in the performance indicators of models a_2, b_2 and c_2, only model *a* is analyzed later in detail due to its higher simplicity (translated into a lower number of inputs).

The average quality parameters of the model external performance is presented in Table 7. Each model was tested outside the training stations, in the remaining 29 stations, and the performance indicators were rearranged as follows. A mean value was calculated for each test station corresponding to the performance of the remaining 29 station models there. As observed, the performance trend is in general quite similar to the local performance. When no ancillary data supply is considered, the model accuracy can be improved through the introduction of R_a or J as local inputs, with a decrease around 0.1 in the RMSE (0.3422 in model a_1 vs 0.2420 in b_1 and 0.2418 in c_1, respectively). Further, the consideration of ancillary exogenous inputs also involves an improvement in the model performance, with a decrease in the RMSE ranging between 0.14 and 0.17 in model *a* and between 0.4 and 0.7 in models *b* and *c*, depending on the number of ancillary stations considered. Thus, the accuracy also improves with an increasing number of ancillary stations, but this improvement is not so marked as in the local performance case. So, the performance quality of the models is considerably worse. As in the local case, models a_2, b_2 and c_2 show very similar results, probably for the same reason suggested above. In contrast to the r^2 values of the local performance, an increasing trend is missing in the external performance. Here, the determination coefficients more or less remain constant around 0.87–0.88, clearly lower than in the local performance case. Further, there is not a clear trend in terms of over-/underestimation, attending to the MBE values. Despite the worsening of the performance trends, it must be pointed out that the individual values used to calculate these means correspond to 29 external models. So, these results can be distorted by the not considering only the most suitable models for each test station. Consequently, it seems more appropriate to assess in each test station only those models trained in the most suitable corresponding training stations [31]. According to the RMSE, the a_2, b_2 and c_2 models providing the optimum local performance do not always fit with those providing the optimum external performance (e.g. optimum model b_2 corresponds to 11 ancillary stations). Nevertheless, the differences are very slight. So, no distinction will be considered between optimum local and external performance and only the local and external performance of the best a_2 model (with 12 ancillary stations) will be analyzed later in detail.

The selection procedure of the optimum network architecture is represented in Fig. 6, corresponding to the training station 30 and model a_2 with 12 ancillary inputs. Here, the relationship MSE-number of neurons of the hidden layer is analyzed. Moreover, these relationships are depicted for the three defined data sets: the training, the cross-validation and the test sets. The horizontal line represents the lowest value of the MSE referring to the cross-validation, in this case 475.83 $(W/m^2)^2$. Thus, the configuration 1 hidden layer with 15 neurons was selected. These results correspond to the optimum repetition for each architecture: the repetition with the lowest MSE in the cross-validation set. A subjective criterion would have lead to the selection of other architectures, seeking for simpler configurations presenting only slightly higher cross-validation and test errors than the current ones. Nevertheless, given the high number of model cases studied, the selection process of the optimum configuration demanded automation.

Accordingly, Table 8 sums up the selected architectures of the models a_1, b_1 and c_1 as well as the optimum a_2, b_2, c_2 (with 12, 12 and 10 ancillary stations, respectively) models in every training station. It seems logical to question the convenience of detecting trends or relationships within the obtained configurations, given the absence of a clear and definitive methodology to deal with the optimum architecture selection in the ANN community. Nonetheless, the average configurations of the models which do not consider ancillary data supply are slightly more complex than their corresponding pairs with ancillary supply, with 3–5 mean neurons more on average, respectively. Likewise, configurations with less than 10 hidden neurons are more frequent in the models which consider exogenous R_s as inputs. The higher network complexity can be due to more complex input–output relationships, when only local records are considered for R_s estimation.

The RMSE values presented in Table 9 allow a detailed analysis of the external performance corresponding to the optimum a_2 model (12 ancillary stations). As aforementioned, a mean external

Table 8
Selected network configurations of optimum models.

Training station code	Optimum number of hidden neurons					
	Model					
	a_1	b_1	c_1	a_2	b_2	c_2
1	8	9	20	8	13	6
2	16	13	11	8	9	16
3	9	12	18	3	11	10
4	10	10	16	3	5	10
5	16	8	13	10	6	12
6	10	20	19	4	3	6
7	14	16	16	20	7	11
8	13	8	13	14	12	6
9	20	17	20	4	9	6
10	18	12	14	17	17	11
11	20	15	20	12	18	5
12	4	12	16	12	6	10
13	6	20	18	11	5	11
14	10	19	11	9	8	3
15	19	7	9	8	11	14
16	16	9	18	7	6	5
17	12	19	20	15	19	7
18	5	12	12	8	19	20
19	11	18	8	8	10	5
20	8	18	10	5	7	20
21	14	16	11	9	12	4
22	19	9	20	13	2	11
23	7	13	19	16	4	20
24	17	16	11	8	8	11
25	10	18	16	19	19	6
26	12	19	9	8	17	11
27	12	7	14	14	19	11
28	17	13	15	17	16	7
29	15	14	20	12	3	6
30	13	6	14	15	12	11
Mean	12.7	13.5	15.0	10.6	10.4	9.7

Please cite this article in press as: Martí P, Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage (2010), doi:10.1016/j.enconman.2010.08.027

Table 9
External performance analysis of model a_2.

Test station code	RMSE (–)					Corresponding training station code	
	Mean	Minimum	Maximum	5th best	Standard deviation	Optimum	Worst
1	0.1794	0.0970	0.3173	0.1437	0.0438	24	15
2	0.1758	0.1215	0.2961	0.1542	0.0337	15	5
3	0.1546	0.1085	0.2589	0.1210	0.0383	16	15
4	0.1847	0.1466	0.3383	0.1559	0.0357	9	15
5	0.1376	0.0918	0.2099	0.0982	0.0363	15	16
6	0.1527	0.0988	0.2840	0.1209	0.0395	8	10
7	0.1912	0.1081	0.3085	0.1615	0.0375	22	25
8	0.1916	0.1350	0.3338	0.1642	0.0382	18	15
9	0.1579	0.0801	0.2520	0.0998	0.0452	7	22
10	0.1708	0.0984	0.2536	0.1245	0.0425	6	19
11	0.1591	0.0882	0.2843	0.1302	0.0392	5	23
12	0.1805	0.1171	0.3171	0.1478	0.0392	6	11
13	0.1804	0.1191	0.3126	0.1447	0.0400	26	29
14	0.2434	0.2178	0.2886	0.2270	0.0154	25	17
15	0.1798	0.1246	0.2307	0.1503	0.0276	30	13
16	0.1602	0.1120	0.2294	0.1285	0.0306	17	3
17	0.1535	0.0900	0.2479	0.1210	0.0358	16	4
18	0.1870	0.0924	0.2767	0.1240	0.0551	19	29
19	0.1519	0.0804	0.2408	0.1058	0.0385	16	6
20	0.1689	0.0870	0.3666	0.1324	0.0501	11	13
21	0.1757	0.1237	0.2516	0.1368	0.0333	27	5
22	0.1784	0.0989	0.3148	0.1517	0.0396	27	29
23	0.1831	0.1172	0.3629	0.1431	0.0447	5	11
24	0.1684	0.0996	0.3615	0.1179	0.0511	9	4
25	0.1761	0.1189	0.3045	0.1450	0.0372	20	22
26	0.1568	0.1025	0.3533	0.1259	0.0452	22	11
27	0.1667	0.1085	0.2862	0.1294	0.0375	5	21
28	0.1723	0.1224	0.2275	0.1369	0.0319	2	23
29	0.1538	0.1203	0.2501	0.1327	0.0248	26	19
30	0.1806	0.1516	0.2349	0.1586	0.0221	19	1
Mean	0.1724	0.1126	0.2865	0.1378	0.0377	–	–

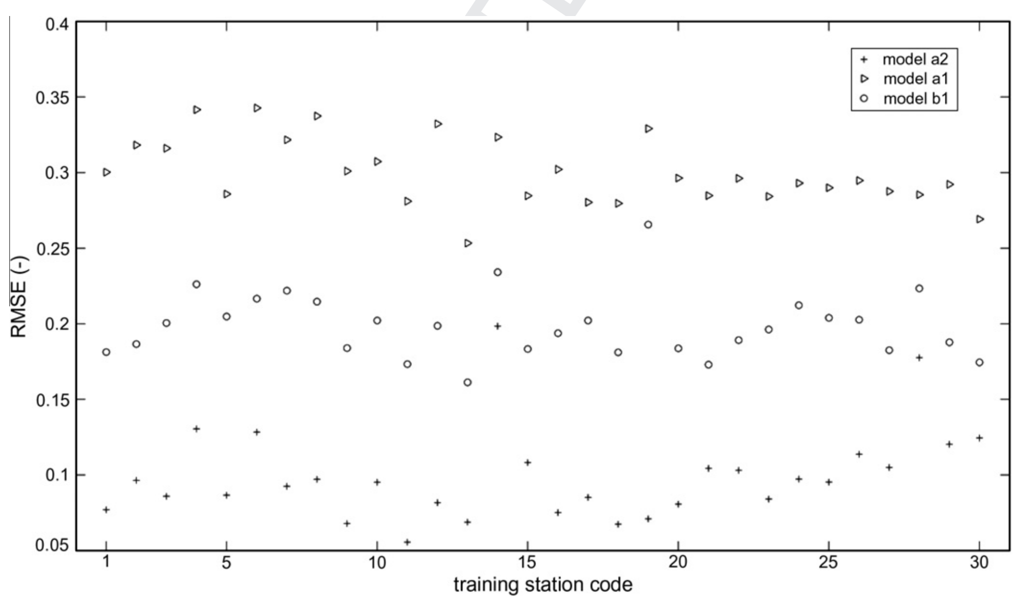

Fig. 7. RMSE values corresponding to the local performance of models a_1, a_2 and b_1 in the studied stations.

performance per station might not be justified, given the heterogeneity associated within the 29 stations considered to provide the mean external performance. So, this table brings together for each station (when considered as test station) the performance achieved: (a) averaging the rest of training station performances (column 2), (b) with the optimum training station for that test station, (c) with the worst training station for that test station and (d) with the fifth best training station for that test station. Comparing the values in columns 2 and 3, remarkable differences can be found between the performances corresponding to the optimum and the worst training stations, as it was foreshadowed. Nevertheless, it might be difficult to select a priori the most appropriate training station because of a probable lack of suitable information. Thus, this table presents a more conservative case, the fifth optimum training station. These results demonstrate that it is not convenient to take into account the complete set of remaining training stations

Please cite this article in press as: Martí P, Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage (2010), doi:10.1016/j.enconman.2010.08.027

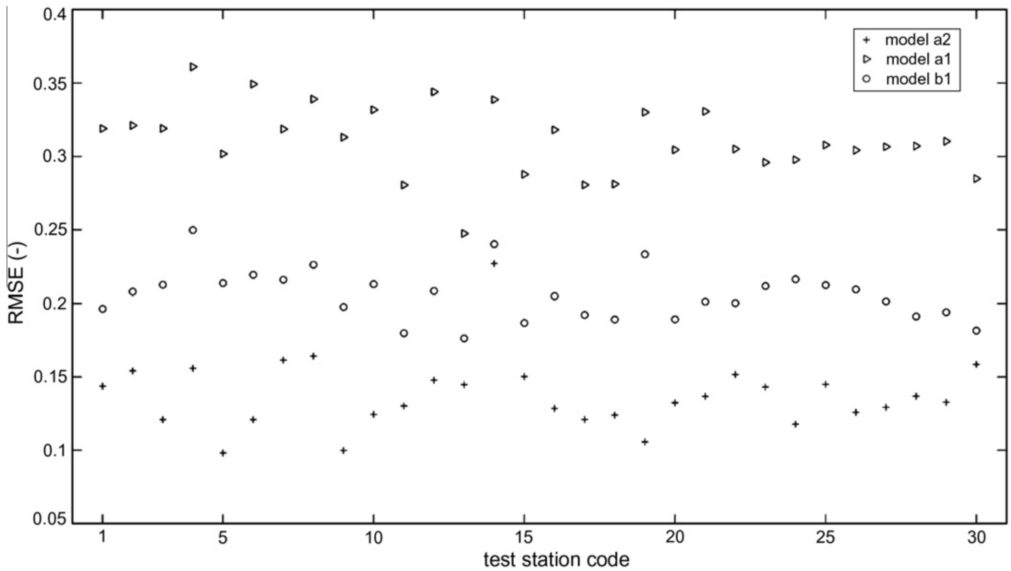

Fig. 8. RMSE values corresponding to the external performance of models a_1, a_2 and b_1 in the studied stations.

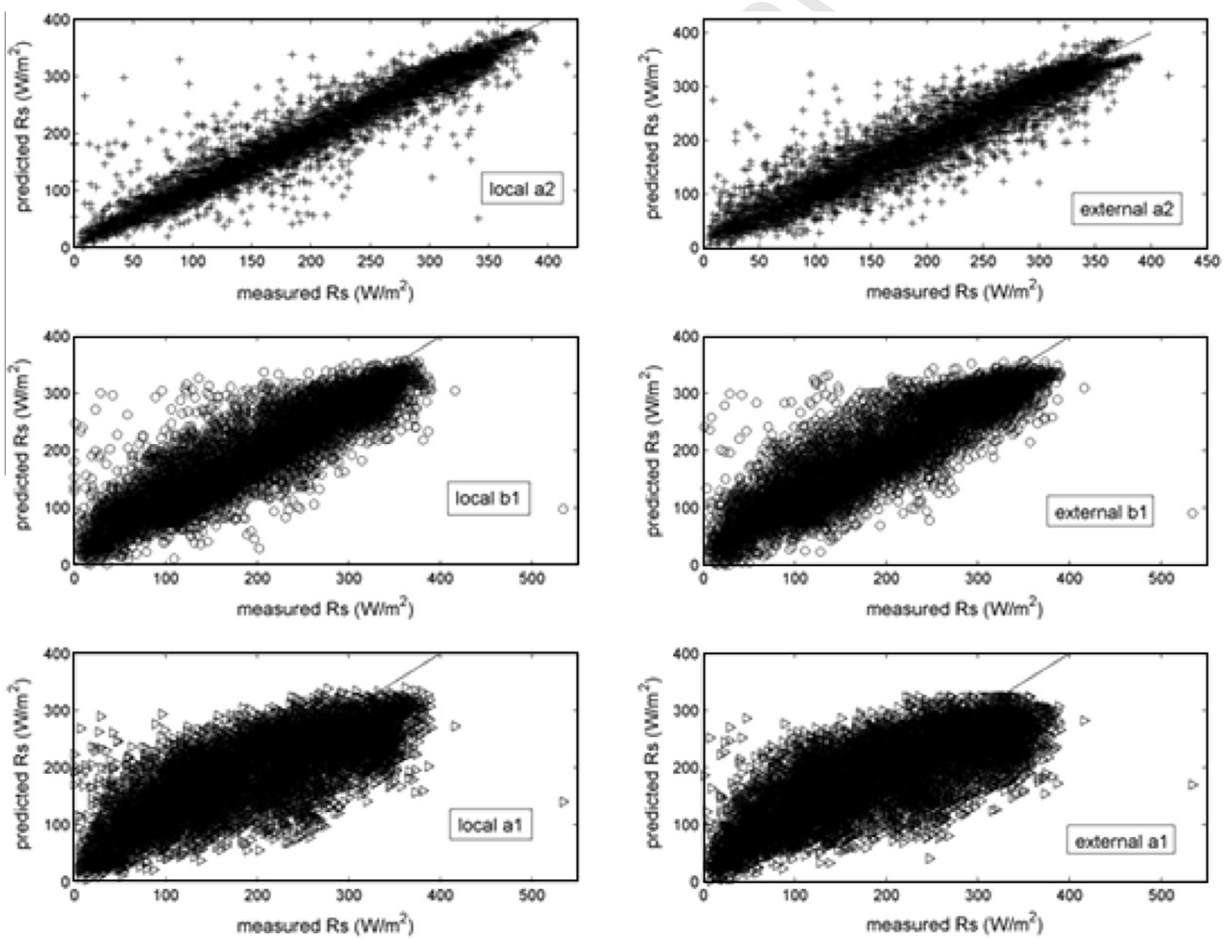

Fig. 9. Scatter plots of models a_1, a_2 and b_1 in the studied stations.

when evaluating the external performance in each test station. The mentioned RMSE fluctuations could derive from the consideration of stations under markedly different solar conditions. A similar study carried out in a smaller area within a homogeneous set of similar stations might not have justified this procedure. Finally, it is difficult to find a clear trend from the analysis of the optimum and worst training stations, although some of them are repeated several times. Hence, when dealing with the external performance of a model, a proper analysis of the test and training station relationships is mandatory to select the most suitable training station.

Therefore, the oscillation ranges of the involved climatic variables of both stations must be considered.

It is important to highlight that 15 of the 29 test stations where each model is performed might correspond to stations previously used as ancillary data suppliers in the training process of that model. Nevertheless, the process can still be considered as external performance in these test stations because the ancillary stations and the corresponding assignment order do not coincide in the training and test stages. Further, the weights established during the training process for a specific exogenous R_s input are assigned in the test stage of this model to the exogenous R_s input of another station, according to the assignment order of Table 4. A detailed analysis of Table 4 and columns 7 and 8 of Table 9 shows that for the model a_2 with 12 secondary stations the test station (column 7 of Table 9) was considered as R_s input supplier for that optimum training station only in 10 models (test stations 4, 7, 10, 11, 13, 14, 19, 21, 23 and 26), whereas the test stations were considered as ancillary suppliers of the worst training stations (column 8 of Table 9) in 7 models (test stations 2, 6, 7, 11, 18, 25 and 27). Despite the satisfactory results, further research should focus on the refinement of the index used in the ancillary station selection. Moreover, the relationships between this index and the weights of the corresponding associated selected ancillary variables with the targets should be analyzed.

Next, Fig. 7 presents the individual RMSE values corresponding to the local estimations obtained in each station with models a_1, b_1 and a_2 with 12 ancillary stations. As observed, model b_1 is always considerably more accurate than model a_1 and model a_2 is always markedly more accurate than the latter two. So, the local performance of temperature-based ANN models for solar radiation estimation can be improved through the consideration of exogenous R_s inputs obtained in similar stations or, if these are not available, through the consideration of R_a or J as inputs, which do not require experimental measurement. With the exception of stations 14 and 28 (with RMSE values near 0.2), the proposed model presents average RMSE values around 0.093, reaching minimum values of 0.0555 (station 11). Hence, the proposed ancillary data supply procedure is decisive to improve the performance of temperature-based ANN models when they are tested in the training station. Given that local R_s records are used as targets to carry out the training process, the usefulness of the developed models in the training stations is limited to emergency or infilling models to be considered when breakdowns take place in the data acquisition system or when alternative more precise models cannot be applied, because there are not enough climatic measurements for their application.

Likewise, Fig. 8 shows a comparison of the individual RMSE values corresponding to the estimations obtained in each test station with the aforementioned models, when they are trained outside. Instead of selecting the optimum model (training station) for each test station, a more conservative criterion was adopted for this comparison and so, these results correspond to the fifth best training station per test station. Here, very similar trends to those of the local performance can be noticed, with a clear worsening in the model accuracy. Neglecting station 14, the proposed model allows R_s estimations with average RMSE around 0.125 in stations where only temperature records are available, reaching minimum error values of 0.1. Nevertheless, the estimations might be more accurate with a more appropriate selection of the training station.

Finally, Fig. 9 shows the scatter plots corresponding to the local and fifth best external performances of the models a_1, a_2 and b_1 in the 30 stations. In comparison to the models a_1 and b_1, the graphics of model a_2 present around 1500–2000 points less due to the matrix homogenization process associated to the consideration of ancillary exogenous inputs. The improvement associated to the proposed model is remarkable. As can be seen, model b_1 is considerably more accurate than a_1 and model a_2 is markedly more accurate than the latter two. Here, a_2 estimations present clearly lower dispersion.

As pointed out in the introduction, the accuracy of the temperature-based models for R_s estimation depends highly on the temperature range of the test location. So, further research should focus on the improvement rates that are to be achieved according to the proposed methodology in areas with different ΔT ranges.

4. Conclusions

This paper describes a new procedure to improve the performance accuracy of temperature-based ANN models for solar radiation estimation through the consideration of exogenous inputs from secondary similar stations, which work as ancillary data suppliers. Thus, this model only demands maximum and minimum temperature records from the studied station. The Gorezynski continentality index is used to select the most appropriate secondary stations.

The accuracy of the model performance improves with an increasing number of ancillary stations. Nonetheless, if the number of ancillary stations considered is too high, the number of training patterns decreases considerably due to the homogenization process established to remove data gaps and it might not be enough to fulfill a proper training. Further, the increase in the number of ancillary stations considered cannot infringe the similarity condition between stations.

Given that local solar radiation records are used as targets to carry out the training process, the usefulness of the developed models in the training stations is limited to emergency or infilling models to be considered when breakdowns take place in the data acquisition system or when alternative more precise models cannot be applied, because there are not enough climatic measurements for their application. On the other hand, when dealing with the external performance of the model, where its application might be of more interest, a careful selection of the most suitable training station is mandatory.

References

[1] Hontoria L, Aguilera J, Zufiria P. Generation of hourly irradiation synthetic series using the neural network multilayer perceptron. Sol Energy 2002;72(5):441–6.
[2] Zekai Ş. Solar energy fundamentals and modeling techniques: atmosphere, environment, climate change and renewable energy. Springer; 2008.
[3] López G, Rubio MA, Martínez M, Batlles FJ. Estimation of hourly global photosynthetically active radiation using artificial neural network models. Agric Forest Meteorol 2001;107:279–91.
[4] Reddy KS, Ranjan M. Solar resource estimation using artificial neural networks and comparison with other correlation models. Energy Convers Manage 2003;44:2519–30.
[5] Meza F, Varas E. Estimation of mean monthly solar global radiation as a function of temperature. Agric Forest Meteorol 2000;100:231–41.
[6] Tymvios FS, Jacovides CP, Michaelides SC, Scouteli C. Comparative study of Ångström's and artificial neural networks' methodologies in estimating global solar radiation. Sol Energy 2005;78:752–62.
[7] Mohandes M, Rehman S, Halawani TO. Estimation of global solar radiation using artificial neural networks. Renew Energy 1998;14(1–4):179–84.
[8] Rehman S, Mohandes M. Artificial neural network estimation of global solar radiation using air temperature and relative humidity. Energy Policy 2008;36:571–6.
[9] Bosch JL, López G, Batlles FJ. Daily solar irradiation estimation over a mountainous area using artificial neural networks. Renew Energy 2008;33:1622–8.
[10] Azadeh A, Maghsoudi A, Sohrabkhani S. An integrated artificial neural networks approach for predicting global radiation. Energy Convers Manage 2009;50:1497–505.
[11] Liu X, Mei X, Li Y, Wang Q, Jensen JR, Zhang Y, et al. Evaluation of temperature-based global solar radiation models in China. Agric Forest Meteorol 2009;149:1433–46.
[12] Hunt LA, Kuchar L, Swanton CJ. Estimation of solar radiation for use in crop modelling. Agric Forest Meteorol 1998;91:293–300.

[13] Abraha MG, Savage MJ. Comparison of estimates of daily solar radiation from air temperature range for application in crop simulations. Agric Forest Meteorol 2008;148:401–16.
[14] Fadare DA. Modelling of solar energy potential in Nigeria using an artificial neural network model. Appl Energy 2009;86:1410–22.
[15] Haykin S. Neural networks. A comprehensive foundation. New Jersey: Prentice Hall International Inc.; 1999.
[16] Patterson DW, editor. Artificial neural networks. Theory and applications. Singapore: Prentice Hall; 1996.
[17] Galvão CO, Valença MJS, Vieira VPPB, Diniz LS, Lacerda EGM, Carvalho ACPLF, et al. Intelligent systems: applications to hydric resources and environmental sciences. Federal University of Rio Grande do Sul/Brazilian Association of Water Resources. Brazil: Porto Alegre, Rio Grande do Sul; 1999.
[18] Al-Alawi SM, Al-Hinai HA. An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation. Renew Energy 1998;14(1–4):199–204.
[19] Mohandes M, Balghonaim A, Kassas M, Rehman S, Halawani TO. Use of radial basis functions for estimating monthly mean daily solar radiation. Sol Energy 2000;68(2):161–8.
[20] Sfetsos A, Coonick AH. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Sol Energy 2000;68(2):169–78.
[21] Dorvlo ASS, Jervase JA, Al-Lawati A. Solar radiation estimation using artificial neural networks. Appl Energy 2002;71:307–19.
[22] Kalogirou S, Michaelides S, Tymvios F. Prediction of maximum solar radiation using artificial neural networks. In: Proceedings of the WREC VII, Germany; 2002.
[23] Sözen A, Arcaklioğlu E, Özalp M. Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. Energy Convers Manage 2004;45:3033–52.
[24] Lam JC, Wan KKW, Yang L. Solar radiation modelling using ANNs for different climates in China. Energy Convers Manage 2008;49:1080–90.
[25] 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.
[26] López G, Batlles FJ, Tovar-Pescador J. Selection of input parameters to model direct solar irradiance by using artificial neural networks. Energy 2005;30:1675–84.
[27] Hontoria L, Aguilera J, Zufiria P. An application of the multilayer perceptron: solar radiation maps in Spain. Sol Energy 2005;79:523–30.
[28] Hargreaves GH, Samani ZA. Estimating potential evapotranspiration. J Irrig Drain Eng 1982;108(3):225–30.
[29] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric Forest Meteorol 1984;31:159–66.
[30] Allen RG. Self-calibrating method for estimating solar radiation from air temperature. J Hydrol Eng 1997;2(2):56–67.
[31] Martí P, Gasque M. Ancillary data supply strategies for improvement of temperature-based ET_o ANN models. Agric Water Manage 2010;97(7):939–55.
[32] Font I. Climatic atlas of Spain. Madrid: National Institute of Meteorology; 1983.
[33] Hagan MT, Menhaj MB. Training multilayer networks with the Marquardt algorithm. IEEE Trans Neural Network 1994;5(6):989–93.
[34] Hagan MT, Delmuth H, Beale M. Neural network design. Boston: MA PWS Publishing; 1996.
[35] MATLAB users' manual version 7.4.0 (R2007a), The MathWorks Inc., Natick, Mass; 2007.
[36] Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press; 1997.