Arabian Journal of Geosciences (2020) 13:431
https://doi.org/10.1007/s12517-020-05437-0

ORIGINAL PAPER

New input selection procedure for machine learning methods in estimating daily global solar radiation

Seyed Mostafa Biazar[1] · Vahid Rahmani[2] · Mohammad Isazadeh[1] · Ozgur Kisi[3] · Yagob Dinpashoh[1]

Received: 9 December 2019 / Accepted: 16 May 2020

Abstract

Selection of optimal model inputs is a challenging issue particularly for non-linear and dynamic systems. In this study, a new input selection method, procrustes analysis (PA), was implemented and compared with gamma test (GT) for estimating daily global solar radiation (Rs). The PA and GT were applied for modeling with the non-linear models of artificial neural networks (ANNs) and support vector machines (SVMs). Goodness-of-fit of the models was evaluated by the coefficient of correlation (CC), root-mean-square error (RMSE), and Nash-Sutcliffe model efficiency coefficient (NS). The uncertainty of the model outputs was determined using 95PPU% (p-factor) and d-factor. In this study, we used maximum wind speed, mean wind speed, maximum temperature, minimum temperature, mean temperature, maximum sea surface pressure, minimum sea surface pressure, mean sea surface pressure, mean vapor pressure, total rainfall, maximum cloudiness, mean cloudiness, maximum humidity, minimum humidity, mean humidity, sunshine hours, evaporation, mean dew point temperature, mean wet point temperature, maximum air pressure, minimum air pressure, mean air pressure, and mean vapor saturation as input variables. Maximum and mean temperature; maximum wind speed; maximum, minimum, and mean sea surface pressure; maximum, minimum, and mean air pressure; mean vapor pressure; mean cloudiness; mean humidity; sunshine hours; mean dew point temperature; mean wet point temperature; and mean vapor saturation pressure were identified as significant input variables by GT in five or more of the eight studied stations. Also, mean air pressure, mean cloudiness, and mean temperature were identified as significant input variables for Rs modeling by the PA method for more than four stations. Results indicated that although ANN-GT and SVM-GT showed better goodness-of-fit metrics, ANN-PA and SVM-PA had lower uncertainties for estimating Rs. According to the obtained results, almost all models showed that the higher the bandwidth (95PPU or P-factor), the greater the d-factor, and the lower the bandwidth, the lower the d-factor, SVM-PA has the lowest uncertainty among the four models. So, it can be seen that the lowest bandwidth also belonged to the SVM-PA model for Kiashahr with a P-factor of 0.8% and a d-factor of 0.06, although the Aliabad-E-Katoul had the lowest d-factor of 0.017 and a p-factor of 1%. The highest d-factor belonged to the ANN-GT model for a Bandar-E-Torkman with a d-factor of 0.817 and a p-factor of 76%. One reason for the high uncertainty in this model might be due to the number of input variables selected by the GT. Lower uncertainty is a major scale for choosing the optimal model for solving a given problem, suggesting results of the SVM-PA model with lower uncertainty are more reliable.

Keywords Procrustes analysis · Gamma test · Solar radiation · Input selection · Non-linear systems

Introduction

Solar radiation (Rs) is the main source of energy for the earth (Yang et al. 2006; Jamil and Akhtar 2017) and plays an important role in climate change (Donatelli et al. 2006; Mercado et al. 2009; Perdigão et al. 2017). Rs has an important role in atmospheric, oceanic, and hydrologic sciences (Bray and Han 2004; Zhang et al. 2017; Li et al. 2019a, b). However, direct measurements of Rs are not broadly obtainable across the world (Liu and Wang et al. 2019). Lack of photovoltaic radiation information is quite prevalent even in developed

Responsible Editor: Zhihua Zhang

✉ Ozgur Kisi
ozgur.kisi@iliauni.edu.ge

[1] Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran

[2] Department of Biological and Agricultural Engineering, Kansas State University, Manhattan, KS 66503, USA

[3] School of Technology, Ilia State University, Tbilisi, Georgia

Published online: 04 June 2020

countries, such as the United States (Richardson 1985; Hook and McClendon 1992) and Canada (Jong and Stewart 1993; Fan et al. 2019). Different general approaches exist for estimating Rs, including empirical, artificial neural network (ANN), and satellite-based methods (Xu et al. 2016; Jahani and Mohammadi 2018). In this study, the ANN and support vector machine (SVM), two common methods used in previous studies, were used for estimating Rs (Lopez et al. 2005; Şenkal and Kuleli 2009; Benghanem et al. 2009; Mohandes 2012; Zajaczkowski et al. 2013; Chen and Li 2014; Zang et al. 2019; Adnan et al. 2019a, b). The amount of required accuracy and available meteorological data are important factors in selecting a method. However, one of the main challenges in estimating a variable using the mentioned models is the type and number of variables that should be used as inputs of the model. Although model performance is generally improved by inclusion of more information, not all inputs enhance prediction accuracy. This is because additional input data may cause the model to be overfitted. The overfitted model also performs well in training but has a very poor performance in testing. It is therefore important to know which inputs are effective in modeling and which are not (Ahmadi et al. 2009).

Remesan et al. (2008) used the gamma test (GT) for determining effective inputs for estimating Rs using local linear regression (LLR) and ANN in the Brue catchment in southwest England. Results showed that out of six input variables, four had more impact on Rs estimation including precipitation, daily maximum temperature, daily minimum temperature, and extraterrestrial radiation. The LLR model showed better performance than the ANN based on root-mean-square error (RMSE) and coefficient of determination (R^2) values using the same input variables for each model. Ahmadi et al. (2009) applied GT, entropy theory (ET), Akaike's information criterion (AIC), and Bayesian information criterion (BIC) to determine effective input data for Rs estimation in the Brue catchment in England. The GT reduced the number of the significant input variables to three (horizontal extraterrestrial radiation, air bulb temperature, wet bulb temperature) out of five. Results showed that the ANN-GT outperformed the other models based on the RMSE metric. Guermoui et al. (2018) used the Gaussian process regression (GPR) for estimating the Rs in Algeria and found sunshine hours, minimum air temperature, and relative humidity as the best combination.

For estimating Rs in Iran, Shamshirband et al. (2016) estimated the diffuse Rs using a coupled SVM-wavelet transform model (SVM-WT) in Kerman, Iran. They used the daily diffuse fraction (cloudiness index) correlated with the daily clearness index as input variables and found that SVM-WT was more efficient method for estimating Rs. Vakili et al. (2017) presented an ANN-based model for modeling Rs utilizing particle matter, wind speed, relative humidity, and temperature as input variables in Tehran, Iran. They concluded that including particle matter to climatological parameters improved the accuracy of Rs predictions. Jahani and Mohammadi et al. (2018) comprised the performance and suitability of different types of models for estimating Rs based on sunshine hours and diurnal air temperature in Iran. These models consist of empirical, classic ANN, and ANN models combined with a genetic algorithm (ANN-GA). Their results showed that ANN-GA had better performance than other methods. Samadianfard et al. (2019) modeled daily global Rs using data-driven techniques and empirical equations in semi-arid regions in Iran. They applied gene expression programming (GEP), model trees (MT), support vector regression (SVR), adaptive neuro fuzzy inference system (ANFIS), and several empirical equations. Several meteorological variables were used in the study including minimum and maximum temperature, sunshine hours, relative humidity, maximum sunshine hours, cloudiness, day of year, and extra-terrestrial radiation as model inputs. Their results showed that the MT model had more precision than the other methods for estimating Rs.

The mentioned methods have been used extensively along with other methods to estimate different variables worldwide. For example, Rashidi et al. (2016) forecasted sediment load with SVM-GT with two kernels: the radial basis function (RBF) and the polynomial and conventional regression method in Korkorsar, Iran. Four parameters were selected by GT as input variables for sediment-load modeling out of the nine studied variables. Variables included flow discharge and suspended-sediment discharge with and without lag. Suspended-sediment discharge with one and two lag, flow discharge without lag, and with one and two lags, and were selected by the conventional regression method as the best input variables. Their results indicated that the SVM-GT with an RBF kernel estimated sediment load with more accuracy than other methods based on statistical criteria of R^2, RMSE, Nash-Sutcliffe efficiency coefficient (NS), mean absolute error (MAE), and mean relative error (MRE). Tian et al. (2016) applied data-driven models with GT to predict groundwater dynamics in northern China. Their models included the power function model (PFM), back-propagation artificial neural network (BPANN), and SVM. Precipitation, groundwater exploitation, irrigation groundwater use, industrial groundwater use, crop yield, gross domestic product, primary industry output, secondary industry output, tertiary industry output, population, and urbanization were chosen by GT as significance input variables. Their findings indicated that the SVM model performed the best in terms of RMSE, NS, and R^2.

Mohammadi et al. (2018) investigated the performance of combined GT and ANFIS for estimating reservoir sediment in Sistan, Iran. Percentages of sand, clay, silt, organic carbon cation exchange capacity, and total Pb were selected by GT as input variables. Results showed that ANFIS-GT had good performance based on RMSE and R^2. Singh et al. (2018) investigated rainfall-runoff modeling in a hilly watershed in

the Ramaganga River Basin in Uttarakhand, India. They used the co-adaptive neuro fuzzy inference system (CANFIS) and multi-layer perceptron (MLP). To select the best combination of input variables, they used GT. The best combination parameters included rainfall without lag and with one lag, and runoff with one lag and two lags. Their results indicated better accuracy for the CANFIS compared to the MLP in terms of RMSE. Seifi and Riahi (2018) investigated daily reference evapotranspiration estimation using a hybrid GT-least-square support vector machine (LSSVM), and ANN-GT and ANFIS-GT models in Iran. Minimum and maximum air temperature and wind speed were selected by GT as the most important input parameters. Their results indicated that the LSSVM model provided better accuracy than the ANFIS and ANN models when similar meteorological input variables were used. Notton et al. (2019) carried out a research about Rs estimation with ANN method. They investigated their results with respect to normalized root mean square error (nRMSE). The results indicated that the ANN method performance has been good. Antonopoulos et al. (2019) conducted a research about Rs estimation in Greece. They used ANN, multi-linear regression (MLR), and empirical method (Hargreaves). In the study, they used daily meteorological measurements of air temperature, radiation, humidity, and wind velocity. The results showed that the ANN and MLR methods had better performance than the empirical model. Rabehi et al. (2020) estimated Rs with ANN, boosted decision tree (BDT), and a new combination of these models with linear regression (LR) in the south of Algeria. The achieved results showed outperformance of the ANN model.

A regular objective in heuristic multivariate analysis is to determine a subset of the variables that transmits the main aspect of the whole instance. Analysis of a well-known multivariate data set indicates that methods currently existed for deciding variables in principal component analysis (PCA) may not lead to a suitable subset (Krzanowski 1987). In this study, a new selection method, procrustes analysis (PA), is suggested and shown to lead to a better selection of input variables for estimating Rs. A limited number of studies have used PA to optimize the number and type of model inputs. Dinpashoh et al. (2004) investigated the variable selection for precipitation regionalization in Iran. They used the PA method for selecting the best combination of input variables for their model. They found 12 (out of 57) variables with stronger impact on precipitation regionalization. Nam et al. (2015) used PA for delineating rainfall in South Korea and found 33 (out of 42) rainfall-related and geographical variables for best results. Zinchenko et al. (2019) made the analysis of relations between communities of hydrobionts in saline rivers by multidimensional block ordination. They used PA to isolate stable associations of taxa typical for particular biotopes with homogeneous environmental conditions. Despite as mentioned, many studies conducted on Rs estimation, the literature on comparison of different methods for choosing inputs for estimating Rs is limited. In the present investigation, for the first time, GT and PA methods have been investigated with the help of non-linear models of ANN and SVM. It should be mentioned that according to the authors' research, the limited researches existed about Rs estimation in shoreline region especially in northern Iran.

The main objectives of this study are to (a) identify the best combination of input variables for estimating Rs using GT and PA and (b) estimate Rs values using ANN and SVM and determine the model with lower uncertainty.

Materials and method

This section provides brief information about the case study, used data and methods, and analysis implemented. Figure 1 illustrates the main steps of the experiments followed in the presented study.

Study area

Two provinces of Guilan and Golestan in the northern coastal region of Iran were studied. Guilan and Golestan are located between 48° 32′ N and 50° 36′ N and 54° N and 56° N, respectively (Fig. 1). Guilan is predominantly covered by tall forest trees and has a humid subtropical climate with hot and humid summers and mild winters (Biazar and Ferdosi 2020; Isazadeh et al. 2017; Dinpashoh et al. 2019; Biazar et al. 2019). Golestan has mild weather and temperatures most of the year and is geographically divided into two parts of plains and mountains. Three meteorological stations at Rasht, Lahijan, and Kiashar from Guilan and five stations at Aliabad-E-Katoul, Bandar-E-Torkaman, Gonbad-E-Kavus, Maravehtappeh, and Gorgan from Golestan were analyzed (Fig. 2 and Table 1).

Data

The meteorological data were downloaded from the Iran Meteorological Organization (http://www.irimo.ir/far/). Daily values of maximum and mean wind speed (m/s); minimum, maximum, and mean temperature (°C); minimum, maximum, and mean sea surface pressure (pa); minimum, maximum, and mean air pressure (pa); mean vapor pressure (pa); total rainfall (mm); maximum and mean cloudiness (octa); minimum, maximum, and mean humidity (%); total sunshine hours (hrs); mean evaporation (mm); mean dew point temperature (°C); mean wet point temperature (°C); and mean vapor saturation pressure (pa) were analyzed for the eight stations. The data quality was carefully checked by drawing the time series of each station and visually inspecting the outliers in the data.

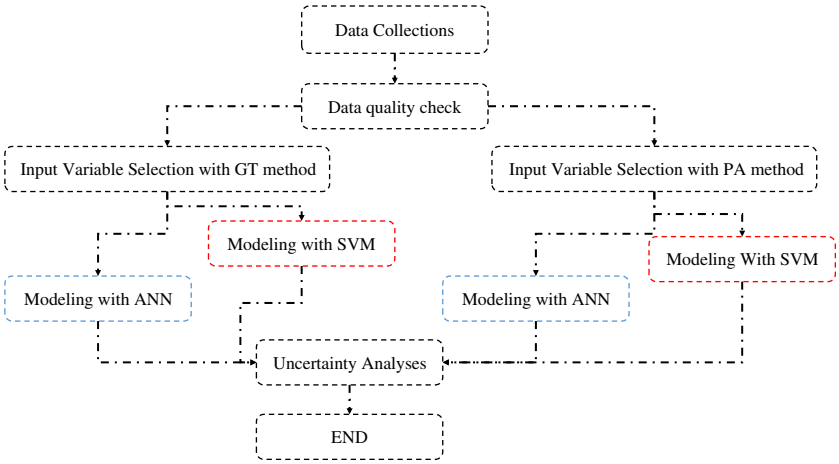


Fig. 1 Flowchart introducing the main steps and processes of the experiments

Artificial neural network

Artificial neural networks (ANNs) are considered as one of the data processing methods, including input/output processing and generally one or more hidden layers as their main components. Even though there are various neural network architectures, about 90% of them are the feed-forward type (Coulibaly et al. 2000; Li et al. 2019a, 2019b; Aghelpour et al. 2019). A feed-forward ANN can have one or more hidden layers which their nodes are called as hidden units or nodes. The ANN can solve high-ordered statistical problems by including hidden layers. A typical three-layer ANN structure is shown in Fig. 1. In the figure, Wij indicates the weights between input layer i and hidden layer j while Wjk are the weights of the network between hidden layer j and output layer k. The ANN can have several outputs; however, in the presented study, one output was utilized. The perceptron, the very basic form of an artificial neural network, is a binary classifier and can be described using the following equation:

$$g(z) = \begin{cases} 1 \; if \; z = w \times X + b > 0 \\ 0 \; otherwise \end{cases} \quad (1)$$

where $g(z)$ is the Heaviside step function (i.e., a limited activation function), w is the weight vector, b is the perceptron parameter deviation, and X is the input vector. Multilayer perceptron (MLP) networks learn to simulate the treatment of a wrapped, nonlinear system through learning algorithms

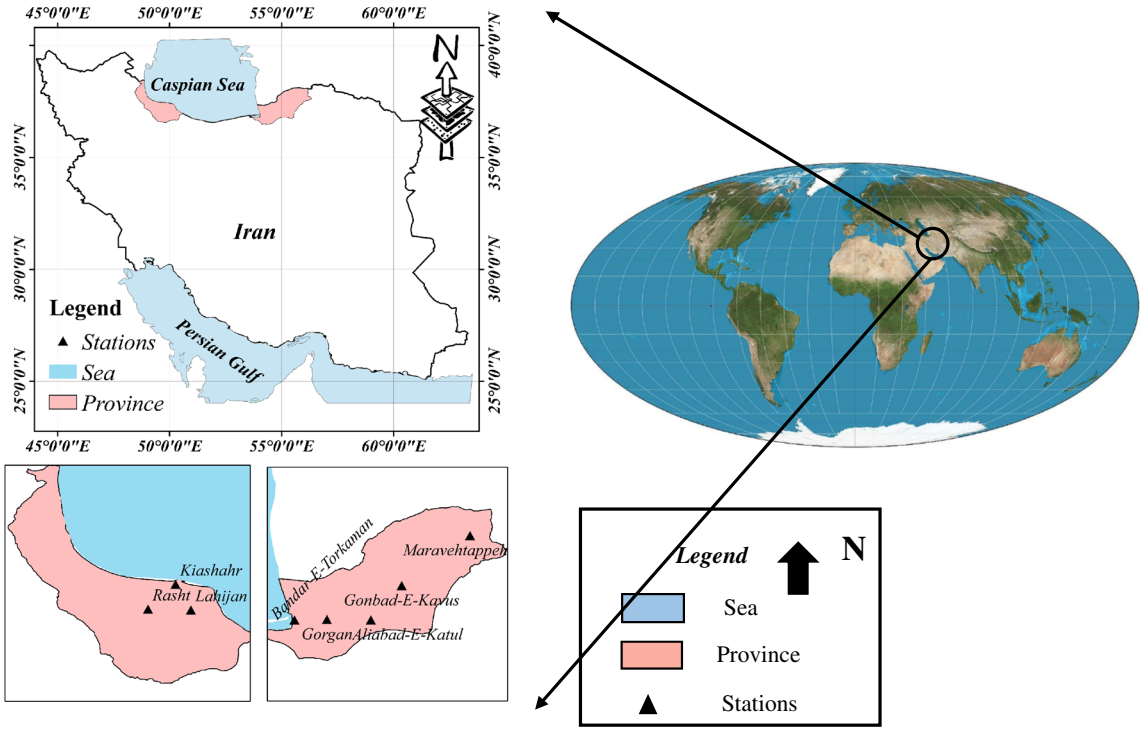


Fig. 2 Location of the study area and selected stations

Table 1 Characteristics of stations used in this study

Station	Data period	Latitude	Longitude	Mean values of Rs (J/cm^2/day)	Elevation (m)	Number of data points
Kiashahr	7/18/2015–11/2/2017	37° 42′	49° 88′	1179	−22	252
Lahijan	1/14/2016–11/21/2018	37° 19′	50° 01′	1425	34	312
Rasht	4/18/2016–11/21/2018	37° 20′	49° 64′	945	25	283
Aliabad-E-Katul	3/10/2014–11/21/2018	36° 9′	54° 88′	986	184	514
Bandar-E-Torkaman	3/2/2016–11/21/2018	36° 9′	54° 06′	1837	0	297
Gonbad-E-Kavus	9/9/2015–11/21/2018	37° 27′	55° 21′	815	37	350
Gorgan	7/30/2012–11/21/2018	36° 91′	54° 41′	1531	0	686
Maravehtappeh	2/25/2014–11/21/2018	37° 8′	55° 94′	1655	460	518

and observed data. There are different kinds of learning algorithms, the most famous of which is back-propagation (Rumelhart et al. 1988), delta-bar-delta (Jacobs 1988), quick-prop (Fombellida and Destiné 1992), conjugate gradient (Charalambous 1992), and Levenberg-Marquardt (Hagan and Menhaj 1994; Deo et al. 2018; Ashrafzadeh et al. 2019). These learning algorithms are generally used to detect the optimal set of MLP model parameters. This study used the MLP model along with learning algorithm Levenberg-Marquardt learning algorithm. A total of 1 to 20 neurons in the hidden layer are employed to evaluate the effects of network structure on its performance in RS simulation. The "trial and error" method has been used to obtain the optimal neuron. The sigmoid tangent function is applied to map information from the input layer to the hidden layer and from the hidden layer to the output layer (Kisi and Yildirim 2005; Lagos-Avid and Bonilla 2017; Naganna et al. 2019). Moreover, this study used 70% of the data as the training set and the remaining 30% as the test set Fig 3.

Support vector machine

Support vector machine (SVM) is a well-known method utilized for classification, pattern recognition, regression, and performance approximation (Vapnik et al. 1997; Dibike et al. 2001). SVMs can be classified into two classes for a set of vectors (Vapnik et al. 1997). They are constructed based on a hyperplane as $w.\ X+b=0$ that divides a set of n-dimensional vectors $X_i \in R^n$ into two groups. This optimized hyperplane is located farthest from the support vectors and closest to the data points of every class. Finding w is equivalent to solving a second-class programming problem. To overcome this problem, an exchange parameter ($c > 0$) must be determined. To classify linearly inseparable vectors, various kernel functions including degree-d polynomial, radial basis, or hyperbolic tangent can be utilized for multidimensional mappers viewed in a higher-dimensional space (Ashrafzadeh et al. 2018; Choubin et al. 2019; Ashrafzadeh et al. 2020). The following radial basis function was used:

$$k\left(X_i, X_j\right) = \exp\left(-\gamma\left\|X_i - X_j\right\|^2\right) \quad (2)$$

where $\gamma > 0$ is the parameter of the kernel and X_i, X_j represent feature vectors in some input space. The nonlinear regression version of SVMs is written as

$$y = \sum_{i=1}^{m}\left(g_i - g_i^*\right)k\left(X_i, X_j\right) + b \quad (3)$$

where m indicates total number of input data; g_i and g_i^* are the Lagrange multipliers for upper and lower constraints, respectively; and k denotes the kernel function employed to map the n-dimensional input vectors. The same data set and input vectors were used for calibrating and validating MLP and developing SVM models.

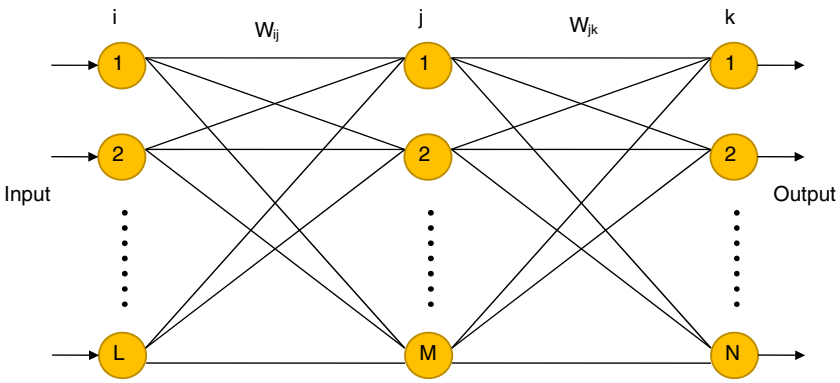


Fig. 3 A three-layer ANN architecture

Gamma test

The gamma test (GT) is utilized to see the relationship between inputs and outputs in numerical data sets. In GT, it is assumed that the observed datasets are represented as follows:

$$\{(x_i - y_i), 1 \le i \le M\} \quad (4)$$

where x_i is the input observation vector, y_i is the observed output, and M is the total number of observations. The basic idea is quite distinct from earlier attempts with non-linear analysis (Durrant 2001; Evans and Jones 2002). When running GT, the normalization option was used to increase processing speed (Ahmadi et al. 2009; Parsaie et al. 2017).

Principal component analysis

Principal component analysis (PCA), perhaps the oldest multivariate method, was introduced by Pierson for the first time in 1901 and later on by Hotelling (1933). The main idea in PCA is to reduce the dimensions of the data that feature a relatively high correlation in terms of modulus. To reduce dimensions, data are converted into new variables that are PCs or indicators independent from one another. The process is carried out in such a way that the first several indicators incorporate a large part of the variation in the entire data. The changes are larger in the first indicator than the second one, and in the second indicator more than the third, and so on (Johnson and Hanson 1995). Instead of making direct use of the input variables, they can be converted to PCs that can be subsequently applied as the throughputs. Input variables are offered with the least loss of the principal components (Johnson and Hanson 1995).

Procrustes analysis

In order to display the general form of a series of multivariate data, it is necessary to calculate the number of dominant indicators (k) and the series of selected PC values for these indicators. Indicator values are linear combinations of all the variables (p). If p is very large, the number of PCs and eigenvalues will be very large, and the interpretation of the indicators will encounter problems. One of the solutions to this problem is selection of a reasonable number of variables (e.g., q) among all variables in such a way that the information from all candidate variables be in the context of all selected variables. In this study, the method suggested by Krzanowski (1987) was used for this purpose. It is necessary to select a very low number of elements in a subsystem in contrast to the reference collection ($q \ll p$). Then, the number of these elements should be equal or larger than the number of selected principal components ($q \ge k$). Selection of the variables is carried out based on optimization through the minimization

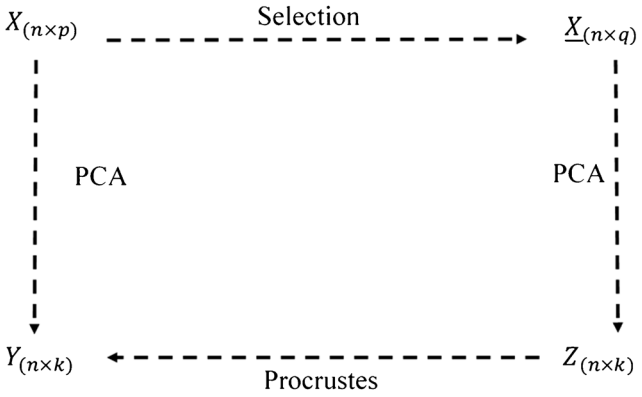


Fig. 4 Cyclone of variable selection using the PA method (Dinpashoh et al. 2004)

of the objective function designated by M^2 (Eq. (4)). Figure 4 illustrates the general stages of PA.

In Fig. 4 and Eq. (5), matrix X is the standardized data, matrix $X_{\_}$ is the standardized data obtained from stepwise elimination of the variables, matrix Y is computed based on

Table 2 Summary of the GT results for estimating daily Rs in Guilan. Values in bold associate with the input parameters that improved Rs estimation significantly

No.	Eliminate	Rasht Gamma	Lahijan Gamma	Kiashahr Gamma
1	–	0.0783	0.0379	0.0279
2	Max. wind speed[a]	0.0690	**0.0425**	0.0267
3	Mean wind speed	**0.0834**	0.0373	0.0236
4	Max. temperature	0.0766	**0.0385**	**0.0283**
5	Min. temperature	0.0718	0.0368	0.0275
6	Mean temperature	**0.0806**	**0.0390**	**0.0290**
7	Max. sea surface pressure	0.0748	**0.0380**	**0.0281**
8	Min. sea surface pressure	0.0768	**0.0382**	**0.0294**
9	Mean sea surface pressure	0.0747	0.0374	**0.0285**
10	Mean vapor pressure	0.0756	**0.0382**	**0.0288**
11	Total rainfall	**0.0843**	0.0367	**0.0287**
12	Max. cloudiness	0.0695	**0.0384**	0.0213
13	Mean cloudiness	**0.0793**	**0.0382**	0.0247
14	Max. humidity	**0.0814**	**0.0456**	0.0273
15	Min. humidity	0.0642	0.0347	0.0225
16	Mean humidity	0.0739	**0.0425**	**0.0281**
17	Sunshine hours	0.0760	**0.0485**	**0.0402**
18	Evaporation	0.0737	**0.0419**	0.0273
19	Mean dew point temperature	0.0779	0.0372	**0.0283**
20	Mean wet point temperature	**0.0817**	**0.0379**	**0.0287**
21	Max. air pressure	0.0738	**0.0380**	**0.0287**
22	Min. air pressure	0.0719	**0.0389**	**0.0285**
23	Mean air pressure	0.0750	0.0373	**0.0289**
24	Mean vapor saturation	**0.0793**	**0.0392**	0.0275

[a] All parameters are daily

Table 3 Summary of the GT results for estimating daily Rs in Golestan. Values in bold associate with the input parameters that improve Rs estimation significantly

No.	Eliminate	Maravetappeh Gamma	Gorgan Gamma	Gonbad-E-Kavus Gamma	Bandar-E-Torkaman Gamma	Aliabad-E-Katoul Gamma
1	–	0.0291	0.0953	0.0352	0.0673	0.0389
2	Max. wind speed[a]	**0.0297**	**0.0972**	0.0306	**0.0924**	**0.0447**
3	Mean wind speed	**0.0313**	**0.1117**	0.0278	0.0373	0.0358
4	Max. temperature	**0.0313**	0.0953	**0.0406**	**0.0698**	0.0375
5	Min. temperature	0.0285	**0.0963**	0.0326	0.0592	**0.0439**
6	Mean temperature	**0.0309**	**0.1030**	**0.0365**	**0.0787**	**0.0435**
7	Max. sea surface pressure	**0.0298**	**0.1275**	0.0343	0.0564	**0.0440**
8	Min. sea surface pressure	**0.0301**	**0.1224**	0.0330	**0.0763**	**0.0417**
9	Mean sea surface pressure	**0.0296**	**0.1085**	0.0333	**0.0680**	**0.0441**
10	Mean vapor pressure	**0.0305**	**0.0987**	0.0307	**0.0804**	0.0381
11	Total rainfall	0.0281	**0.1321**	0.0345	**0.0757**	0.0377
12	Max. cloudiness	0.0286	0.0886	0.0320	**0.0825**	0.0368
13	Mean cloudiness	0.0245	**0.0959**	**0.0366**	**0.0688**	0.0354
14	Max. humidity	0.0232	**0.1342**	0.0336	**0.0848**	0.0364
15	Min. humidity	0.0290	**0.1071**	**0.0385**	**0.0841**	**0.0429**
16	Mean humidity	0.0264	**0.1384**	**0.0384**	**0.0732**	**0.0404**
17	Sunshine hours	**0.0429**	**0.1335**	**0.0510**	**0.0852**	**0.0513**
18	Evaporation	0.0284	**0.0986**	0.0338	0.0575	**0.0474**
19	Mean dew point temperature	**0.0345**	**0.1046**	0.0291	**0.0788**	**0.0454**
20	Mean wet point temperature	**0.0332**	**0.1140**	0.0348	0.0793	**0.0453**
21	Max. air pressure	**0.0292**	**0.1353**	0.0342	0.0538	**0.0442**
22	Min. air pressure	0.0283	**0.1321**	0.0317	**0.0696**	**0.0478**
23	Mean air pressure	**0.0300**	**0.1096**	0.0346	**0.0682**	**0.0416**
24	Mean vapor saturation	**0.0300**	**0.1057**	**0.0359**	**0.0803**	0.0379

[a] All parameters are daily

all of the candidate variables and commonly called the true configuration, and matrix Z is based on the selected variables and commonly called the approximate configuration of the data. PA evaluates the differences between the sums of the square of the corresponding points in these two arrangements. The more precisely the variables are selected, the greater the similarity of the two arrangements and the lesser the differences between the approximate and real arrangements. The result of the PA is the calculated difference between the sums of the square of the two approximate and real arrangements that can be obtained from the following:

$$M^2 = Trace\left\{YY' + ZZ' - 2ZQ'Y'\right\} \quad (5)$$

Table 4 Optimal input parameters selected based on PA for each station

Province	Station	Input parameters[a]		
Guilan	Kiashahr	Mean air pressure	Mean vapor pressure	Mean cloudiness
	Lahijan	Mean temperature	Mean cloudiness	–
	Rasht	Mean air pressure	Mean cloudiness	Mean dew point temperature
Golestan	Aliabad-E-Katoul	Mean air pressure	Mean cloudiness	Mean vapor saturation pressure
	Bandar-E-Torkaman	Mean temperature	Mean air pressure	Mean cloudiness
	Gonbad-E-Kavus	Mean sea surface pressure	Mean cloudiness	Mean vapor saturation pressure
	Gorgan	Mean temperature	Min. humidity	–
	Maravehtappeh	Min. temperature	Mean air pressure	Mean humidity

[a] All parameters are daily

where Q is obtained according to the following:

$$Q = VU' \tag{6}$$

where matrices U and V are obtained from the singular value decomposition (SVD) of the matrix $Z'Y$ with $k \times k$ dimensions using the following mathematical form:

$$Z'Y = U\sum V' \tag{7}$$

where $UU' = I_k$, $V'V = VV' = I_k$ and $\sum = diag(\sigma_1, \sigma_2, \ldots, \sigma_k)$. I_k is the identity matrix with dimension $k \times k$. M^2 was determined for each of the subsets pertaining to the set of variables with at least q members. The best subset of the variable reference set was selected in such a way that the lowest amount of M^2 can be scored for that subset (Krzanowski 1987).

Error measures

Three metrics were applied to assess the goodness-of-fit of the estimations including coefficient of correlation (CC), root-mean-square error (RMSE), and Nash-Sutcliffe model efficiency coefficient (NS). One of the most popular evaluation indices is the Nash Sutcliffe Index, whose range varies from 1 to negative infinity. The intervals of 0.75–1, 0.36–0.75, and less than 0.36 for this index in a simulation show very good, satisfactory, and poor performance, respectively (Nash and Sutcliffe 1970; Ashrafzadeh et al. 2018; Nazari-Sharabian et al. 2019; Adnan et al. 2019c).

$$CC = \sqrt{\frac{\left(\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)^2\right)}{\left(\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2\right)}} \tag{8}$$

$$RMSE = \sqrt{\frac{\left(\sum_{i=1}^{N}\left(x_i - \overline{y}\right)^2\right)}{N}} \tag{9}$$

$$NS = 1 - \sqrt{\frac{\left(\sum_{i=1}^{N}\left(x_i - \overline{y}\right)^2\right)}{\left(\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2\right)}} \tag{10}$$

where x_i is the observed value, $\overline{x}$ is the mean of observed values, y_i is the estimated value, $\overline{y}$ is the mean of estimated values, and N is the number of observations.

Uncertainty analysis

P-factor (95PPU%) and d-factor coefficients were proposed to quantify the power of calibration and uncertainty analyses (Abbaspour et al. 2007). Equation (11) was used to determine the average width of the band (d-factor) index:

$$d\text{-}factor = \frac{\overline{\mathrm{d}}\,\mathrm{x}}{\sigma \mathrm{x}} \tag{11}$$

where σx is the standard deviation of the observed data and $\overline{d}\,x$ is the average width of the confidence interval that can be achieved by

$$\overline{\mathrm{d}}\,\mathrm{x} = \frac{1}{\mathrm{k}} \sum_{\mathrm{t}=1}^{\mathrm{K}} \left(\mathrm{X_U} - \mathrm{X_L}\right) \tag{12}$$

where $t = 1, \ldots, K$ is the number of observed data, X_U is the 97.5th percentile of model output, and X_L is the 2.5th percentile of model output.

$$\text{Bracketed by 95PPU} = \frac{1}{\mathrm{k}} \text{count}\left(\mathrm{j} | \mathrm{X_L^l} \le \mathrm{X_{reg}^l} \le \mathrm{X_U^l}\right) \times 100 \tag{13}$$

Where l is the item number from one to k, X_{reg}^l is the observed value on day l, and j is the counter parameter of the number of observed values placed on the 95% prediction uncertainties (95PPU) band. If all values are within the confidence band of uncertainty, they are then bracketed by 95PPU = 100 (Abbaspour et al. 2007; Isazadeh et al. 2017; Nazari-Sharabian et al. 2020; Khaledian et al. 2020).

Results and discussion

Gamma test

Various combinations of input parameters were identified using GT in different stations for improving Rs estimation. Maximum and mean temperature; maximum wind speed; maximum, minimum, and mean sea surface pressure; maximum, minimum, and mean air pressure; mean vapor pressure; mean cloudiness; mean humidity; sunshine hours; mean dew point temperature; mean wet point temperature; and mean vapor saturation pressure were identified as significant input variables in five or more of the eight studied stations (Tables 2 and 3). Mean wind speed was eliminated from all stations using the GT method except for Gorgan, Maravetappeh, and Rasht. Minimum temperature was eliminated for all three stations in Guilan. Maximum cloudiness was eliminated for all stations except Bandar-E-Torkaman and Lahijan. Evaporation was eliminated for all stations except Gorgan, Aliabad-E-Katoul, and Lahijan (Tables 2 and 3).

Table 5 Training results and uncertainty analysis of ANN and SVM models

Province	Station	Model	CC	RMSE (J/cm^2/day)	NS	d-factor	95PPU%	Model structure[a]
Guilan	Kiashahr	ANN-GT	0.940	233	0.884	0.506	57	(14-17-1)
		SVM-GT	0.930	251	0.864	0.047	7	RBF
		ANN-PA	0.885	318	0.783	0.374	35	(3-10-1)
		SVM-PA	0.889	312	0.789	0.030	2	RBF
	Lahijan	ANN-GT	0.888	405	0.787	0.688	67	(16-6-1)
		SVM-GT	0.932	321	0.790	0.092	11	RBF
		ANN-PA	0.802	525	0.644	0.314	23	(2-3-1)
		SVM-PA	0.803	523	0.645	0.04	3	RBF
	Rasht	ANN-GT	0.877	269	0.770	0.448	41	(7-5-1)
		SVM-GT	0.874	273	0.763	0.254	14	RBF
		ANN-PA	0.855	293	0.729	0.405	35	(3-14-1)
		SVM-PA	0.854	293	0.728	0.118	10	RBF
Golestan	Aliabad-E-Katoul	ANN-GT	0.927	212	0.859	0.414	54	(15-8-1)
		SVM-GT	0.930	207	0.865	0.109	14	RBF
		ANN-PA	0.860	288	0.739	0.241	19	(3-3-1)
		SVM-PA	0.862	286	0.743	0.016	1	RBF
	Gonbad-E-Kavus	ANN-GT	0.912	186	0.830	0.38	38	(7-19-1)
		SVM-GT	0.909	188	0.830	0.099	8	RBF
		ANN-PA	0.840	244	0.706	0.323	26	(3-5-1)
		SVM-PA	0.843	242	0.706	0.05	5	RBF
	Bandar-E-Torkman	ANN-GT	0.836	530	0.697	0.7	77	(17-20-1)
		SVM-GT	0.830	538	0.688	0.112	13	RBF
		ANN-PA	0.752	635	0.566	0.34	26	(3-2-1)
		SVM-PA	0.764	621	0.584	0.098	8	RBF
	Gorgan	ANN-GT	0.708	638	0.501	0.425	45	(21-3-1)
		SVM-GT	0.732	616	0.535	0.118	11	RBF
		ANN-PA	0.646	689	0.418	0.326	27	(2-11-1)
		SVM-PA	0.647	688	0.419	0.113	9	RBF
	Maravetappeh	ANN-GT	0.876	397	0.767	0.526	46	(14-4-1)
		SVM-GT	0.867	409	0.754	0.072	7	RBF
		ANN-PA	0.815	476	0.665	0.303	23	(3-8-1)
		SVM-PA	0.817	473	0.663	0.02	2	RBF

Several of the identified input parameters in this study were also found significant in previous studies. Remesan et al. (2008) applied GR for estimating Rs and found minimum, maximum, and mean temperature, extra-terrestrial radiation, and mean wind speed as significant input parameters in their model. In the study conducted by Ahmadi et al. (2009), horizontal extraterrestrial radiation, air bulb temperature, and wet bulb temperature were selected as significant estimators for Rs using GT. Mohandes et al. (1998) used latitude, longitude, and sunshine hours for Rs estimation in Saudi Arabia. In Turkey, Sözen et al. (2004) estimated Rs with latitude, longitude, altitude, month, sunshine hours, and temperature as input variables for the MLP model. Azadeh et al. (2009) estimated Rs with maximum and minimum relative humidity, vapor pressure, wind speed, total precipitation, and sunshine hours in Iran. Assi et al. (2013) used maximum temperature, wind speed, sunshine hours, and relative humidity for estimating Rs in the United Arab Emirates. Also, Vakili et al. (2017) used maximum and minimum temperature, relative humidity, and wind speed for estimating Rs in Iran. Marzo et al. (2017) used daily extreme temperatures and extraterrestrial radiation for Rs modeling in desert areas.

Procrustes analysis

In the PA method, the PCA method was used to determine effective parameters for Rs estimation. The number of PCA for each station was determined based on eigenvalues higher than one. Results indicated that effective PCA in Aliabad-E-Katoul, Bandar-E-Torkman, Gonbad-E-Kavus, Gorgan, Kiashahr, Lahijan, Maravetappeh, and Rasht was three, three,

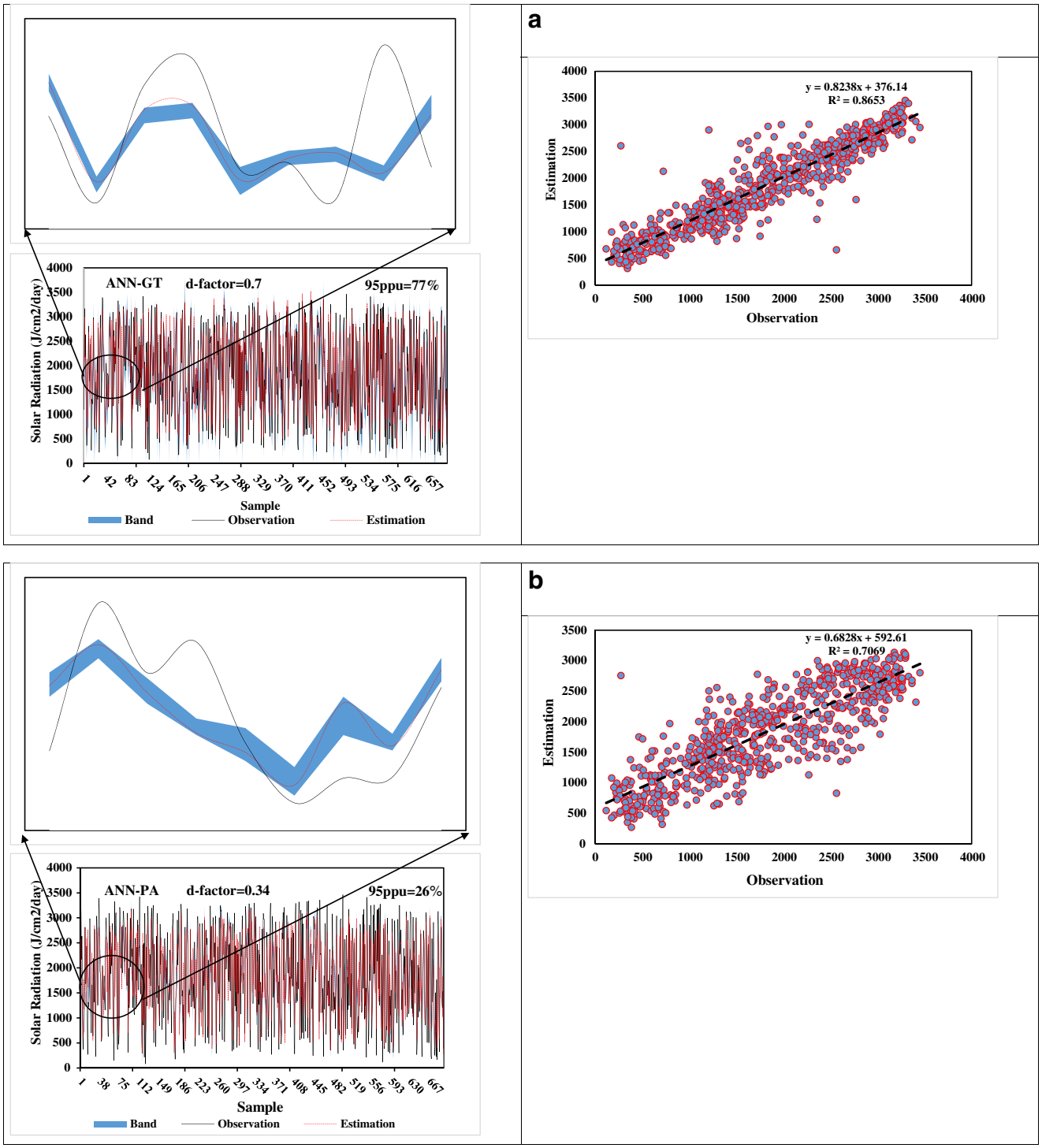


Fig. 5 Uncertainty of ANN-GT (**a**), ANN-PA (**b**), SVM-GT (**c**), and SVM-PA (**d**) for estimating Rs in Bandar-E-Torkman for training period

three, two, three, two, three, and three, respectively. Then, PCA numbers were analyzed by the PA method to determine effective inputs in each station according to the M^2 index among the input parameters and number of PCA (Table 4).

Mean air pressure, mean cloudiness, and mean temperature were identified as significance input variables for Rs modeling by the PA method for four or more stations (Table 4). Mean vapor pressure and mean sea surface pressure were selected just for the Kiashar and Gonbad-E-Kavus stations, respectively. Minimum and mean humidity were the main input variables in Gorgan and Maravetappeh, respectively. In addition to mean vapor saturation, pressure was identified as a significance input variable for two stations of Aliabad-E-Katoul and Gonbad-E-Kavus (Table 4).

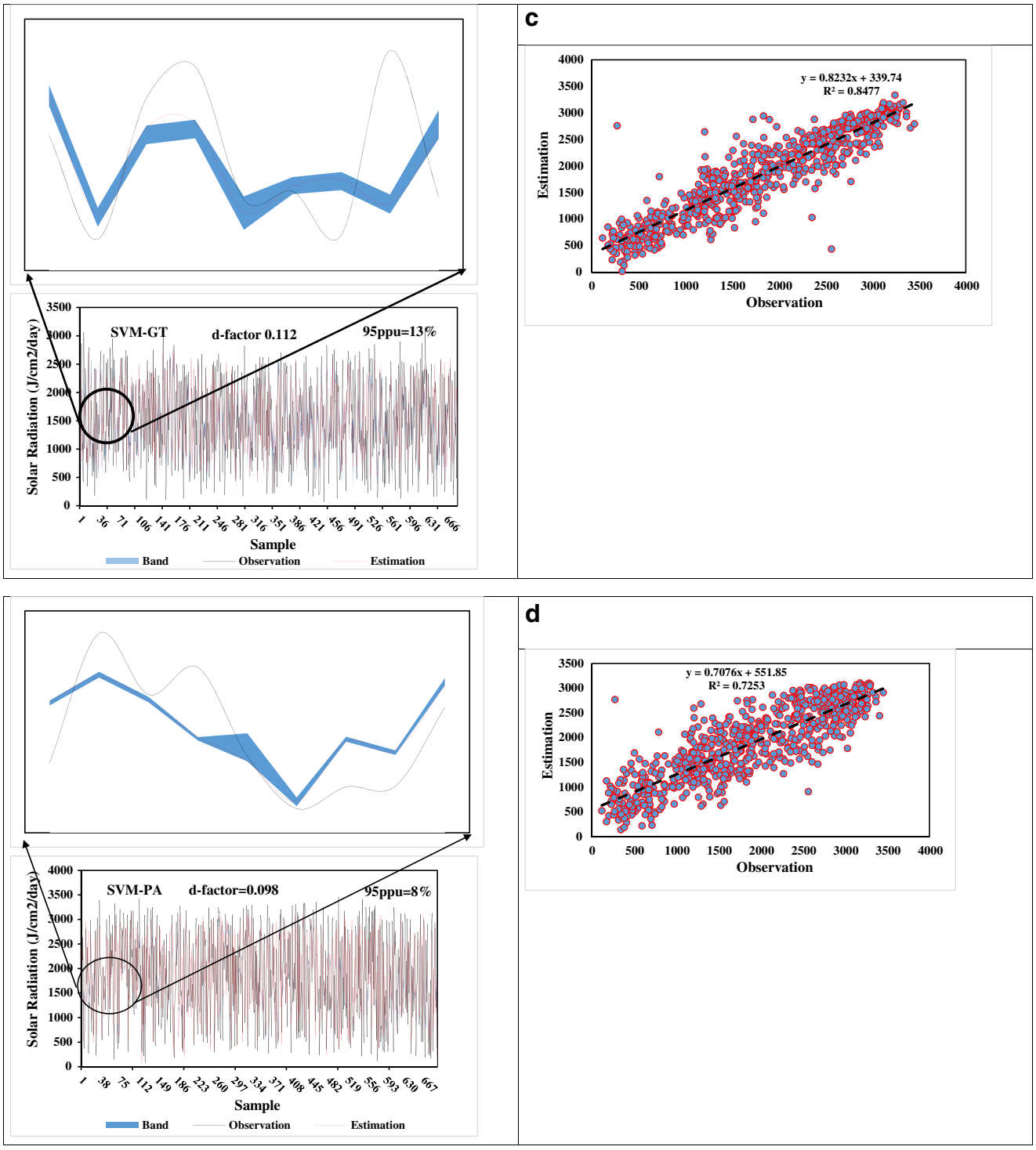


Fig. 5 (continued)

Modeling results

Table 5 shows the results of Rs modeling for training part. In the table, metric indices, such as RMSE, CC, and NS values, were used for comparison of GT and PA-based methods. It is apparent that the ANN-GT performance had better than the other models in all stations (except Aliabad-E-Katoul). This result was obtained, based on the performance measures (RMSE, CC, and NS). As known, one of the uncertainties of statistical and intelligent models is the use of iterations in calculations. So, we used uncertainty analyses in this study. The coefficients, such as the d-factor and p-factor, can be calculated based on the results of each iteration, each representing part of the uncertainty of each model.

Table 6. Validation results and uncertainty analysis of ANN and SVM models

Province	Station	Model	CC	RMSE (J/cm^2/day)	NS	d-factor	95PPU%	Model structure[a]
Guilan	Kiashahr	ANN-GT	0.904	289	0.816	0.551	49	(14-17-1)
		SVM-GT	0.893	304	0.801	0.049	2	RBF
		ANN-PA	0.891	306	0.794	0.392	31	(3-10-1)
		SVM-PA	0.889	307	0.790	0.030	0.8	RBF
	Lahijan	ANN-GT	0.881	412	0.772	0.759	65	(16-6-1)
		SVM-GT	0.869	432	0.743	0.107	9	RBF
		ANN-PA	0.841	472	0.702	0.326	21	(2-3-1)
		SVM-PA	0.837	477	0.681	0.04	4	RBF
	Rasht	ANN-GT	0.813	367	0.660	0.401	42	(7-5-1)
		SVM-GT	0.810	369	0.643	0.203	15	RBF
		ANN-PA	0.803	373	0.646	0.383	38	(3-14-1)
		SVM-PA	0.802	375	0.639	0.105	12	RBF
Golestan	Aliabad-E-Katoul	ANN-GT	0.805	367	0.642	0.474	53	(15-8-1)
		SVM-GT	0.803	368	0.640	0.116	12	RBF
		ANN-PA	0.875	266	0.764	0.254	23	(3-3-1)
		SVM-PA	0.873	267	0.761	0.017	1	RBF
	Gonbad-E-Kavus	ANN-GT	0.881	206	0.774	0.422	34	(7-19-1)
		SVM-GT	0.876	210	0.771	0.104	9	RBF
		ANN-PA	0.812	253	0.657	0.336	27	(3-5-1)
		SVM-PA	0.811	254	0.657	0.049	4	RBF
	Bandar-E-Torkman	ANN-GT	0.902	390	0.810	0.817	76	(17-20-1)
		SVM-GT	0.898	394	0.808	0.153	15	RBF
		ANN-PA	0.828	502	0.683	0.365	26	(3-2-1)
		SVM-PA	0.830	501	0.685	0.103	10	RBF
	Gorgan	ANN-GT	0.721	560	0.517	0.484	47	(21-3-1)
		SVM-GT	0.722	558	0.519	0.124	12	RBF
		ANN-PA	0.649	614	0.421	0.360	24	(2-11-1)
		SVM-PA	0.647	615	0.416	0.123	7	RBF
	Maravetappeh	ANN-GT	0.837	437	0.699	0.595	44	(14-4-1)
		SVM-GT	0.825	451	0.683	0.080	5	RBF
		ANN-PA	0.776	503	0.602	0.330	21	(3-8-1)
		SVM-PA	0.773	505	0.600	0.021	2	RBF

[a] RBF is the kernel that was applied for SVM and ANN models. For instance, 14-7-1 mean 14 inputs, seven hidden, and one output nodes

Therefore, to determine the uncertainty band after a thousand times iteration in each model, a 95% probability band was obtained for each station. The results in Table 5 indicated that although the model ANN-GT revealed better performance than the other models, however, SVM-PA showed lower uncertainty compared to other models, because the values of the d-factor measure obtained for other models were larger than SVM-PA. Results of uncertainty analysis for different stations are shown in Fig. 5. It should be noted that lower uncertainty is more important than the accuracy for selecting the best combination of input variables for estimating a certain output (Abbaspour et al. 2007; Noori et al. 2011; Ghorbani et al. 2016; Isazadeh et al. 2017).

The best input combinations identified by the GT and PA methods were applied to the ANN and SVM models for validation part (Table 6). RMSE, CC, and NS values for ANN-GT and SVM-GT, compared to the values for ANN-PA and SVM-PA, indicated better performance of the former in seven out of eight stations with the exception of Aliabad-E-Katul. Generally, ANN-GT performed the best followed by SVM-GT, ANN-PA, and SVM-PA in terms of model accuracy. However, d-factor and 95PPU% values indicated lower uncertainty for ANN-PA and SVM-PA at most of the stations (Table 6 and Fig. 6). Results showed that the SVM-PA model has the least uncertainty in all stations. ANN-GT and SVM-GT had wider 95PPU% than the ANN-PA and SVM-PA for

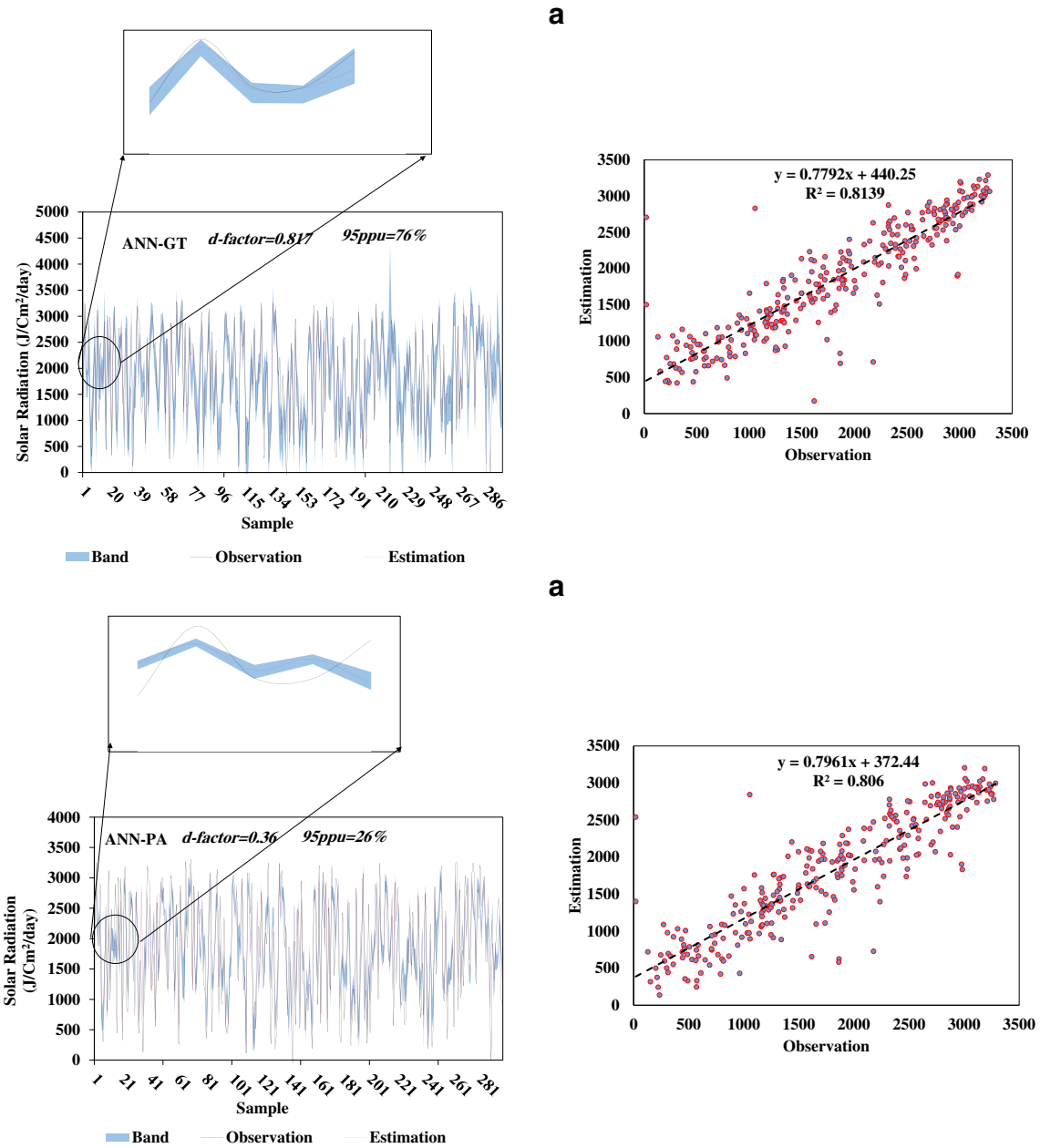
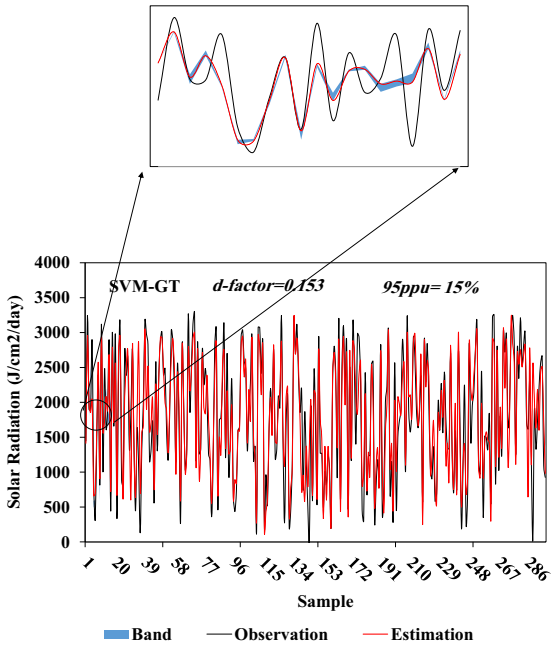
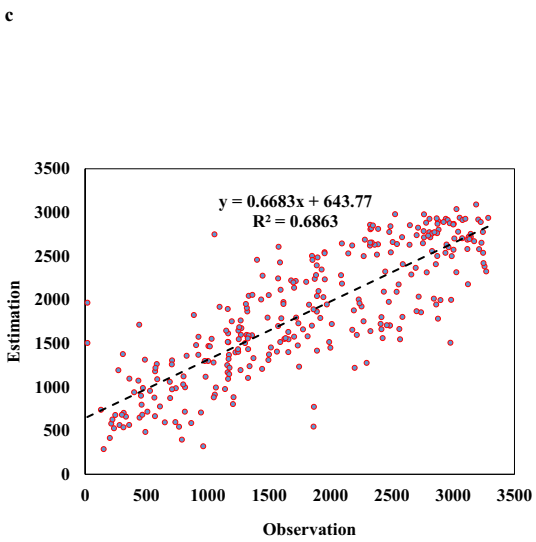
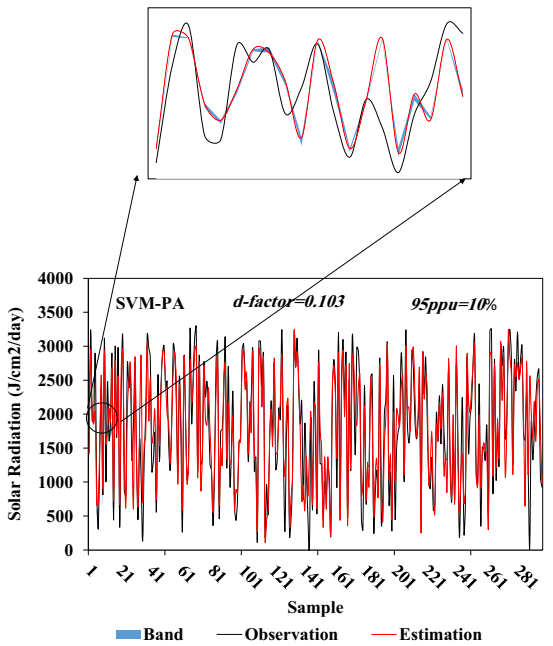


Fig. 6 Uncertainty of ANN-GT (**a**), ANN-PA (**b**), SVM-GT (**c**), and SVM-PA (**d**) for estimating Rs in Bandar-E-Torkman for validation period

most of the stations, suggesting that the d-factor of ANN-GT and SVM-GT is greater than the ANN-PA and SVM-PA. According to results, almost all models showed that the higher the bandwidth (95PPU or P-factor), the greater the d-factor, and the lower the bandwidth, the lower the d-factor, SVM-PA has the lowest uncertainty among the four models. So, it can be seen that the lowest bandwidth also belonged to the SVM-PA model for Kiashahr with a P-factor of 0.8% and a d-factor of 0.06, although the Aliabad-E-Katoul had the lowest d-factor with of 0.017 and with a p-factor of 1%. The highest d-factor belonged to the ANN-GT model for a Bandar-E-Torkman with a d-factor of 0.817 and a p-factor of 76%. One reason for the high uncertainty in this model may be due to the number of input variables selected by the GT (Isazadeh et al. 2017; Khaledian et al. 2020). As a representative example, the uncertainty of ANN and SVM with different scenarios in estimation of Rs is depicted in Figs. 5 and 6; the observed and estimated values of Rs in the training and validation phase, respectively. Generally, results of a model with lower uncertainty are more reliable.







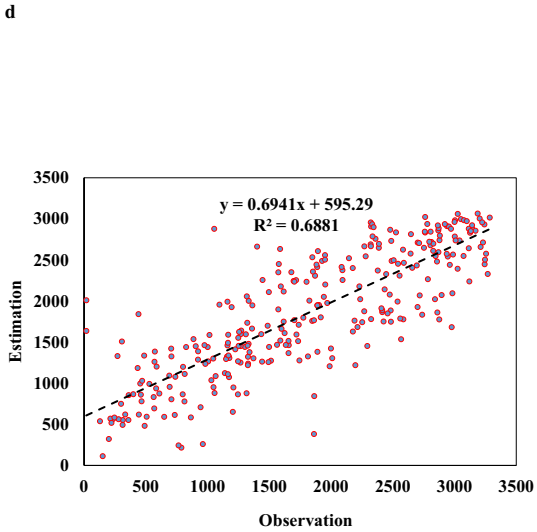


Fig. 6 (continued)

Conclusion

Model input selection is a complicated process, especially for non-linear dynamic systems. Queries on which inputs should be used have been a difficult issue to solve in practice. This study introduced a new input selection method, PA, for estimating Rs. Estimated values from ANN-PA and SVM-PA were compared to ANN-GT and SVM-GT. For each model, goodness-of-fit was evaluated using RMSE, CC, and NS metrics. Uncertainty of each model was determined using 95PPU% and the d-factor. The most important variables identified by GT were the maximum and mean temperature; maximum wind speed; maximum, minimum, and mean sea surface pressure; maximum, minimum, and mean air pressure; mean vapor pressure; mean cloudiness; mean humidity; sunshine hours; mean dew point temperature; mean wet point temperature; and mean vapor saturation pressure in five or more of the eight studied stations. Also, mean air pressure, mean cloudiness, and mean temperature were identified as significance input variables for Rs modeling by the PA

method for more than four stations. Results showed that although the ANN-GT was a quite powerful model and efficient to estimate Rs, SVM-PA generated Rs values with the least uncertainty. Lower uncertainty is more important than accuracy for selecting the best combination of input variables for estimating a certain output. Results of this study showed the power of SVM and PA models for predicting Rs and emphasized the importance of evaluating uncertainty in addition to goodness-of-fit measures for greater reliability of model simulations. According to the results, the PA method was able to demonstrate acceptable performance over the GT methods, with the least number of input variables, in addition to reducing the uncertainty of the models. It can be concluded that by reducing the number of variables as well as selecting the input combination that is a good representative of the effective variables in the Rs estimation, the model complexity can be reduced and a good estimate of Rs can be made. This work will stimulate more researches into input determination methods in model development for other components of climate models and hydrological processes and the authors recommend compared PA method with other input variable selection methods, such as the Entropy theory, by different hybrid models.

Acknowledgments The authors would like to thank the Iran Meteorological Organization for providing data used in this study.

References

Abbaspour KC, Yang J, Maximov I, Siber R, Bogner K, Mieleitner J, Zobrist J, Srinivasan R (2007) Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. J Hydrol 333(2–4):413–430

Adnan RM, Liang Z, Yuan X, Kisi O, Akhlaq M, Li B (2019a) Comparison of LSSVR, M5RT, NF-GP and NF-SC models for hourly wind speed and wind power prediction based on cross-validation. Energies 12(2):329

Adnan RM, Liang Z, Heddam S, Zounemat-Kermani M, Kisi O (2019b) Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. J Hydrol 124371

Adnan RM, Malik A, Kumar A, Parmar KS, Kisi O (2019c) Pan evaporation modeling by three different neuro-fuzzy intelligent systems using climatic inputs. Arab J Geosci 12(20):606

Aghelpour P, Mohammadi B, Biazar SM (2019) Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA. Theoretical Appl Climatol:1–10

Ahmadi A, Han D, Karamouz M, Remesan R (2009) Input data selection for solar radiation estimation. Hydrological Processes: An Int J 23(19):2754–2764

Antonopoulos VZ, Papamichail DM, Aschonitis VG, Antonopoulos AV (2019) Solar radiation estimation methods using ANN and empirical models. Comput Electron Agric 160:160–167

Ashrafzadeh A, Malik A, Jothiprakash V, Ghorbani MA, Biazar SM (2018) Estimation of daily pan evaporation using neural networks and meta-heuristic approaches. ISH J Hydraulic Eng:1–9

Ashrafzadeh A, Ghorbani MA, Biazar SM, Yaseen ZM (2019) Evaporation process modelling over northern Iran: application of an integrative data-intelligence model with the krill herd optimization algorithm. Hydrol Sci J 64(15):1843–1856

Ashrafzadeh, A, Kisi, O, Aghelpor, P, Biazar, S.M, Askarizad, M. (2020). Comparative study of time series models, support vector machines, and GMDH in forecasting long-term evapotranspiration rates in northern Iran, J Irrig Drain Eng, Vol. 146, Issue 6. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001471, 04020010

Assi, A. H., Al-Shamisi, M. H., Hejase, H. A., & Haddad, A. (2013). Prediction of global solar radiation in UAE using artificial neural networks. In 2013 International Conference on Renewable Energy Research and Applications (ICRERA) (pp. 196–200). IEEE.

Azadeh A, Maghsoudi A, Sohrabkhani S (2009) An integrated artificial neural networks approach for predicting global radiation. Energy Convers Manag 50(6):1497–1505

Benghanem M, Mellit A, Alamri SN (2009) ANN-based modelling and estimation of daily global solar radiation data: a case study. Energy Convers Manag 50(7):1644–1655

Biazar SM, Dinpashoh Y, Singh VP (2019) Sensitivity analysis of the reference crop evapotranspiration in a humid region. Environ Sci Pollut Res:1–28

Biazar SM, Ferdosi FB (2020) An investigation on spatial and temporal trends in frost indices in Northern Iran. Theor Appl Climatol. https://doi.org/10.1007/s00704-020-03248-7

Bray M, Han D (2004) Identification of support vector machines for runoff modelling. J Hydroinf 6(4):265–280

Charalambous C (1992) Conjugate gradient algorithm for efficient training of artificial neural networks. IEE Proceedings G (Circuits, Devices and Systems) 139(3):301–310

Chen JL, Li GS (2014) Evaluation of support vector machine for estimation of solar radiation from measured meteorological variables. Theor Appl Climatol 115(3–4):627–638

Choubin B, Moradi E, Golshan M, Adamowski J, Sajedi-Hosseini F, Mosavi A (2019) An Ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci Total Environ 651:2087–2096

Coulibaly P, Anctil F, Bobée B (2000) Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J Hydrol 230(3–4):244–257

Deo RC, Ghorbani MA, Samadianfard S, Maraseni T, Bilgili M, Biazar M (2018) Multi-layer perceptron hybrid model integrated with the firefly optimizer algorithm for windspeed prediction of target site using a limited set of neighboring reference station data. Renew Energy 116:309–323

Dibike YB, Velickov S, Solomatine D, Abbott MB (2001) Model induction with support vector machines: introduction and applications. J Comput Civ Eng 15(3):208–216

Dinpashoh Y, Fakheri-Fard A, Moghaddam M, Jahanbakhsh S, Mirnia M (2004) Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods. J Hydrol 297(1–4):109–123

Dinpashoh, Y., Singh, V. P., Biazar, S. M., & Kavehkar, S. (2019). Impact of climate change on streamflow timing (case study: Guilan Province). Theoretical and Applied Climatology, 1–12.

Donatelli M, Carlini L, Bellocchi G (2006) A software component for estimating solar radiation. Environ Model Softw 21(3):411–416

Durrant PJ (2001) winGamma: a non-linear data analysis and modelling tool with applications to flood prediction. Unpublished PhD thesis, Department of Computer Science, Cardiff University, Wales, UK.

Evans D, Jones AJ (2002) A proof of the Gamma test. Proceedings of the Royal Society of London. Series A: Mathematical. Phys Eng Sci 458(2027):2759–2799

Fan J, Wu L, Zhang F, Cai H, Ma X, Bai H (2019) Evaluation and development of empirical models for estimating daily and monthly mean daily diffuse horizontal solar radiation for different climatic regions of China. Renew Sust Energ Rev 105:168–186

Fombellida M, Destiné J (1992) The extended quickprop. Artif. Neural Networks:973–977. North-Holland. https://doi.org/10.1016/B978-0-444-89488-5.50032-4

Ghorbani MA, Zadeh HA, Isazadeh M, Terzi O (2016) A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction. Environ Earth Sci 75(6): 476

Guermoui M, Gairaa K, Rabehi A, Djafer D, Benkaciali S (2018) Estimation of the daily global solar radiation based on the Gaussian process regression methodology in the Saharan climate. Eur Phys J Plus 133(6):211

Hagan MT, Menhaj MB (1994) Training feedforward networks with the Marquardt algorithm. IEEE Trans Neural Netw 5(6):989–993

Hook JE, McClendon RW (1992) Estimation of solar radiation data missing from long-term meteorological records. Agron J 84(4):739–742

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(6):417–441

Isazadeh M, Biazar SM, Ashrafzadeh A (2017) Support vector machines and feed-forward neural networks for spatial modeling of groundwater qualitative parameters. Environ Earth Sci 76(17):610

Jacobs RA (1988) Increased rates of convergence through learning rate adaptation. Neural Netw 1(4):295–307

Jahani, B., & Mohammadi, B. (2018). A comparison between the application of empirical and ANN methods for estimation of daily global solar radiation in Iran. Theoretical and Applied Climatology, 1–13.

Jamil B, Akhtar N (2017) Estimation of diffuse solar radiation in humid-subtropical climatic region of India: comparison of diffuse fraction and diffusion coefficient models. Energy 131:149–164

Johnson GL, Hanson CL (1995) Topographic and atmospheric influences on precipitation variability over a mountainous watershed. J Appl Meteorol 34(1):68–87

Jong RD, Stewart DW (1993) Estimating global solar radiation from common meteorological observations in western Canada. Can J Plant Sci 73(2):509–518

Khaledian MR, Isazadeh M, Biazar SM, Pham QB (2020) Simulating Caspian Sea surface water level by artificial neural network and support vector machine models. Acta Geophysica:1–11

Kisi Ö, Yildirim G (2005) Discussion of "forecasting of reference evapotranspiration by artificial neural networks" by Slavisa Trajkovic, Branimir Todorovic, and Miomir Stankovic. J Irrig Drain Eng 131(4):390–391

Krzanowski WJ (1987) Selection of variables to preserve multivariate data structure, using principal components. J R Stat Soc: Ser C: Appl Stat 36(1):22–33

Lagos-Avid MP, Bonilla CA (2017) Predicting the particle size distribution of eroded sediment using artificial neural networks. Sci Total Environ 581:833–839

Li DH, Chen W, Li S, Lou S (2019a) Estimation of hourly global solar radiation using multivariate adaptive regression spline (MARS)—a case study of Hong Kong. Energy 186:115857

Li S, Kazemi H, Rockaway TD (2019b) Performance assessment of stormwater GI practices using artificial neural networks. Sci Total Environ 651:2811–2819

Lopez G, Batlles FJ, Tovar-Pescador J (2005) Selection of input parameters to model direct solar irradiance by using artificial neural networks. Energy 30(9):1675–1684

Marzo A, Trigo-Gonzalez M, Alonso-Montesinos J, Martínez-Durbán M, López G, Ferrada P, Batlles FJ (2017) Daily global solar radiation estimation in desert areas using daily extreme temperatures and extraterrestrial radiation. Renew Energy 113:303–311

Mercado LM, Bellouin N, Sitch S, Boucher O, Huntingford C, Wild M, Cox PM (2009) Impact of changes in diffuse radiation on the global land carbon sink. Nature 458(7241):1014–1017

Mohammadi AA, Yousefi M, Soltani J, Ahangar AG, Javan S (2018) Using the combined model of gamma test and neuro-fuzzy system for modeling and estimating lead bonds in reservoir sediments. Environ Sci Pollut Res 25(30):30315–30324

Mohandes MA (2012) Modeling global solar radiation using particle swarm optimization (PSO). Sol Energy 86(11):3137–3145

Mohandes, M., Rehman, S., & Halawani, T. O. (1998). Estimation of global solar radiation using artificial neural networks. Renew Energy, 14(1–4), 179–184, 184.

Naganna SR, Deka PC, Ghorbani MA, Biazar SM, Al-Ansari N, Yaseen ZM (2019) Dew point temperature estimation: application of artificial intelligence model integrated with nature-inspired optimization algorithms. Water 11:742

Nam W, Shin H, Jung Y, Joo K, Heo JH (2015) Delineation of the climatic rainfall regions of South Korea based on a multivariate analysis and regional rainfall frequency analyses. Int J Climatol 35(5):777–793

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—a discussion of principles. J Hydrol 10(3):282–290

Nazari-Sharabian M, Taheriyoun M, Ahmad S, Karakouzian M, Ahmadi A (2019) Water quality modeling of Mahabad Dam watershed–reservoir system under climate change conditions, using SWAT and system dynamics. Water 11(2):394

Nazari-Sharabian M, Taheriyoun M, Karakouzian M (2020) Sensitivity analysis of the DEM resolution and effective parameters of runoff yield in the SWAT model: a case study. J Water Supply Res Technol AQUA 69(1):39–54

Noori R, Karbassi AR, Moghaddamnia A, Han D, Zokaei-Ashtiani MH, Farokhnia A, Gousheh MG (2011) Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction. J Hydrol 401(3–4):177–189

Notton G, Voyant C, Fouilloy A, Duchaud JL, Nivet ML (2019) Some applications of ANN to solar radiation estimation and forecasting for energy applications. Appl Sci 9(1):209

Parsaie A, Azamathulla HM, Haghiabi AH (2017) Physical and numerical modeling of performance of detention dams. J Hydrol 121757

Perdigão J, Salgado R, Magarreiro C, Soares PM, Costa MJ, Dasari HP (2017) An Iberian climatology of solar radiation obtained from WRF regional climate simulations for 1950–2010 period. Atmos Res 198:151–162

Rabehi A, Guermoui M, Lalmi D (2020) Hybrid models for global solar radiation prediction: a case study. Int J Ambient Energy 41(1):31–40

Rashidi S, Vafakhah M, Lafdani EK, Javadi MR (2016) Evaluating the support vector machine for suspended sediment load forecasting based on gamma test. Arab J Geosci 9(11):583

Remesan R, Shamim MA, Han D (2008) Model data selection using gamma test for daily solar radiation estimation. Hydrol Process 22(21):4301–4309

Richardson CW (1985) Weather simulation for crop management models. Trans ASAE 28(5):1602–1606

Rumelhart DE, Hinton GE, Williams RJ (1988) Learning representations by back-propagating errors. Cognitive Model 5(3):1

Samadianfard S, Majnooni-Heris A, Qasem SN, Kisi O, Shamshirband S, Chau KW (2019) Daily global solar radiation modeling using data-driven techniques and empirical equations in a semi-arid climate. Eng Appl Comput Fluid Mech 13(1):142–157

Seifi A, Riahi H (2018) Estimating daily reference evapotranspiration using hybrid gamma test-least square support vector machine, gamma test-ANN, and gamma test-ANFIS models in an arid area of Iran. J Water Clim Change 11:217–240. https://doi.org/10.2166/wcc.2018.003

Şenkal O, Kuleli T (2009) Estimation of solar radiation over Turkey using artificial neural network and satellite data. Appl Energy 86(7–8): 1222–1228

Shamshirband S, Mohammadi K, Khorasanizadeh H, Yee L, Lee M, Petković D, Zalnezhad E (2016) Estimating the diffuse solar

radiation using a coupled support vector machine–wavelet transform model. Renew Sust Energ Rev 56:428–435

Singh A, Malik A, Kumar A, Kisi O (2018) Rainfall-runoff modeling in hilly watershed using heuristic approaches with gamma test. Arab J Geosci 11(11):261

Sözen A, Arcaklioğlu E, Özalp M, Kanit EG (2004) Use of artificial neural networks for mapping of solar potential in Turkey. Appl Energy 77(3):273–286

Tian J, Li C, Liu J, Yu F, Cheng S, Zhao N, Wan Jaafar W (2016) Groundwater depth prediction using data-driven models with the assistance of gamma test. Sustainability 8(11):1076

Vakili M, Sabbagh-Yazdi SR, Khosrojerdi S, Kalhor K (2017) Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data. J Clean Prod 141:1275–1285

Vapnik V, Golowich SE, Smola AJ (1997) Support vector method for function approximation, regression estimation and signal processing. In: Advances in neural information processing systems, pp 281–287. https://doi.org/10.5555/2998981.2999021

Wang L, Lu Y, Zou L, Feng L, Wei J, Qin W, Niu Z (2019) Prediction of diffuse solar radiation based on multiple variables in China. Renew Sust Energ Rev 103:151–216

Xu X, Du H, Zhou G, Mao F, Li P, Fan W, Zhu D (2016) A method for daily global solar radiation estimation from two instantaneous values using MODIS atmospheric products. Energy 111:117–125

Yang K, Koike T, Ye B (2006) Improving estimation of hourly, daily, and monthly solar radiation by importing global data sets. Agric For Meteorol 137(1–2):43–55

Zajaczkowski J, Wong K, Carter J (2013) Improved historical solar radiation gridded data for Australia. Environ Model Softw 49:64–77

Zang H, Cheng L, Ding T, Cheung KW, Wang M, Wei Z, Sun G (2019) Estimation and validation of daily global solar radiation by day of the year-based models for different climates in China. Renew Energy 135:984–1003

Zhang J, Zhao L, Deng S, Xu W, Zhang Y (2017) A critical review of the models used to estimate solar radiation. Renew Sust Energ Rev 70: 314–329

Zinchenko TD, Shitikov VK, Golovatyuk LV, Gusakov VA, Lazareva VI (2019) Analysis of relations between communities of hydrobionts in saline rivers by multidimensional block ordination. Inland Water Biol 12(2):104–110