Stoch Environ Res Risk Assess (2012) 26:59–71
DOI 10.1007/s00477-011-0536-y

ORIGINAL PAPER

Evaluation of efficiency of different estimation methods for missing climatological data

Mahsa Hasanpour Kashani · Yagob Dinpashoh

Published online: 4 November 2011

Abstract Reliable estimation of missing data is an important task for meteorologists, hydrologists and environment protection workers all over the world. In recent years, artificial intelligence techniques have gained enormous interest of many researchers in estimating of missing values. In the current study, we evaluated 11 artificial intelligence and classical techniques to determine the most suitable model for estimating of climatological data in three different climate conditions of Iran. In this case, 5 years (2001–2005) of observed data at target and neighborhood stations were used to estimate missing data of monthly minimum temperature, maximum temperature, mean air temperature, relative humidity, wind speed and precipitation variables. The comparison includes both visual and parametric approaches using such statistic as mean absolute errors, coefficient of efficiency and skill score. In general, it was found that although the artificial intelligence techniques are more complex and time-consuming models in identifying their best structures for optimum estimation, but they outperform the classical methods in estimating missing data in three distinct climate conditions. Moreover, the in-filling done by artificial neural network rivals that by genetic programming and sometimes becomes more satisfactory, especially for precipitation data. The results also indicated that multiple regression analysis method is the suitable method among the classical methods. The results of this research proved the high importance of choosing the best and most precise method in estimating different climatological data in Iran and other arid and semi-arid regions.

Keywords Artificial intelligence and classical techniques · Climatological data · Iran · Missing data

M. Hasanpour Kashani (✉) · Y. Dinpashoh
Department of Water Engineering,
Faculty of Agriculture, University of Tabriz,
Tabriz, Iran
e-mail: mahsakashani2003@yahoo.com

Y. Dinpashoh
e-mail: dinpashoh@yahoo.com

1 Introduction

Estimation of missing data is known as the first stage of most climatological, environmental and hydrological studies. Existence of gaps in the records of data acquisition systems are often attributed to various reasons such as absence of the observer, instrumental failures and communication line breakdown. Developing countries are often faced this problem. Estimation of missing data is more important in mountain and forest regions where meteorological stations are scarce, and the observed data are influenced by topography and the forest microclimate. The techniques of missing data estimation can be grouped in empirical methods, statistical methods and function fitting (Xia et al. 1999). The empirical methods include the simple arithmetic averaging, inverse distance interpolation (ID) (Willmott and Robeson 1995) and ratio and difference technique (Wallis et al. 1991). The statistical methods include multiple regression analysis (REG) (Degaetano et al. 1995), principle component analysis and cluster analysis (Huth and Nemesova 1995), Kriging method (Saborowski and Stock 1994) and optimal interpolation (Bussieres and Hogg 1989). Thin-plate spline technique is one of the function fitting methods which is used to interpolate data (Luo et al. 1998). From the literature review, we find out that blindly using the mentioned methods is common practice in various studies especially

in Iran. This can affect the results of any climatological and hydrological studies.

Mizumura (1985) applied cubic spline (CSP) method to the generation of missing sediment data of the U.S. Geological Survey on Trinity River, California. He found that without measuring sediment concentration in rivers every day, good information for unmeasured data can be found by using the CSP. Eischeid et al. (1995) examined the performance of six methods including empirical and statistical methods for monthly mean temperature and monthly precipitation. They concluded that the REG is the best among others. Xia et al. (1999) estimated the missing data of daily maximum temperature, minimum temperature, mean air temperature, water vapor pressure, wind speed and precipitation using six methods at Bavaria, Germany. They found that the REG method had high ability in estimation of missing data of the study area. Price et al. (2000) interpolated 30-year monthly mean, minimum and maximum temperature and precipitation data from regions in western and eastern Canada using thin-plate smoothing splines (ANUSPLIN) and a statistical method termed Gradient plus Inverse-Distance-Squared. They showed that both interpolators performed best in the eastern region where topographic and climatic gradients are smoother, whereas predicting precipitation in the west was more difficult. In the latter case, ANUSPLIN clearly produced better results for most months. In the last few decades, many types of data-driven techniques have been developed. They reflect the inherently stochastic nature of hydrologic processes and this has led to an increasing interest of researchers in using them. These new machine learning techniques, especially artificial neural networks (ANN) and adaptive neuro-fuzzy inference systems (ANFIS), have been used to solve different hydrological problems during the last decades. Abebe et al. (2000) evaluated the performance of a fuzzy rule-based approach compared with an artificial neural network and a traditional statistical approach reconstruction of missing precipitation events in northern Italy. Results indicated that the fuzzy rule-based model provided solutions with low mean square error (MSE). Srikanthan et al. (2005) used two different approaches to generate daily rainfall data for Sydney and Melborne. These were: (i) a Transition Probability Matrix model and (ii) a nonparametric model. They showed that both approaches preserved most of the variability in rainfall series. Teegavarapu and Chandramouli (2005) developed artificial neural network, kriging and inverse distance weighting method (IDWM) for estimation of missing precipitation data in the state of Kentucky, USA. Results suggested that the conceptual revisions can improve estimation of missing precipitation records by defining better weighting parameters and surrogate measures for distances in the IDWM. Coulibaly and Evora (2007) investigated six different types of ANN namely the multi-layer perceptron (MLP) network and its variations (the time-lagged feed-forward network (TLFN)), the generalized radial basis function (RBF) network, the recurrent neural network (RNN) and its variations (the time delay recurrent neural network (TDRNN)), and the counter propagation fuzzy-neural network (CFNN) along with different optimization methods for infilling missing daily total precipitation records and daily extreme temperature series in Gatineau watershed in northeastern Canada. The experiment results revealed that the MLP can provide the most accurate estimates of the missing precipitation, daily maximum and minimum temperature values. The dynamically driven networks (RBF) appeared fairly suitable only for estimating maximum and minimum temperature. Ustoorikar and Deo (2008) compared genetic programming (GP) and ANN in estimating missing information in hourly significant wave height observations at one of the data buoy stations in the Gulf of Mexico maintained by the US National Data Buoy Center. It was found that the in-filling done by GP rivals that by ANN and many times becomes more satisfactory, especially when the gap lengths are smaller. Kim and Ahn (2009) developed a new spatial daily rainfall model to fill in gaps in a daily rainfall dataset. The model was based on a two-step approach to handle the occurrence and the amount of daily rainfalls separately. They tested four neural network classifiers for a rainfall occurrence processor, and two regression techniques for a rainfall amount processor. The test results revealed that a probabilistic neural network approach is preferred for determining the occurrence of daily rainfalls, and a stepwise regression with a log-transformation is recommended for estimating daily rainfall amounts.

In Iran, according to our best knowledge, most of investigators used blindly a filling method for completing data series in different climatological and hydrological studies. Dastorani et al. (2009) tried to predict the missing data using the normal ratio method (NR), the correlation method, a relevant architecture of ANN as well as ANFIS. According to the results, the ANFIS technique presented a superior ability to predict missing flow data especially in arid land stations of Iran. ANN was also found as an efficient method to predict the missing data in comparison to the traditional approaches.

It seems that there are no significant studies on evaluation of various methods for estimation of climatological missing data in different climates in Iran. The aim of this study is to investigate the capabilities of 11 different traditional and data-driven methods to fill the gaps of hydrometeorological data (monthly minimum temperature, maximum temperature, mean air temperature, relative humidity, wind speed and precipitation) and identify the appropriate method in three distinct climates (Mazandaran,

East Azarbayjan, and Zahedan provinces) of Iran. The eleven candidate methods include arithmetic averaging (AA), inverse distance interpolation (ID), normal ratio method (NR), single best estimator (SIB), multiple regression analysis (REG), UK traditional method (UK), closest station method (CSM), cubic spline (CSP), artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS) and genetic programming (GP).

2 Study area and data description

In this study, 18 weather stations located in different parts of Iran were selected and organized in three groups due to geographical proximity. In each group, one of the stations was selected as a target station. A portion of the recorded data for the target station omitted and it was tried to predict them using the information of nearby stations by different methods. The best method was selected based on three distinct criteria mentioned later in this study. Figure 1 shows the study area and location of the stations across country. Table 1 represents the details of the selected stations. As it can be inferred from Table 1, three climatic regions selected from whole country and for each region six stations indicated. From these six stations a target station selected (indicated by open circles in Fig. 1) and the data of other five nearby stations used for estimation of missing data. The period of 2001–2005 were used in the present study. The climate of each station was determined using the De Martonne aridity index as follows:

$$I = \frac{P}{T + 10} \tag{1}$$

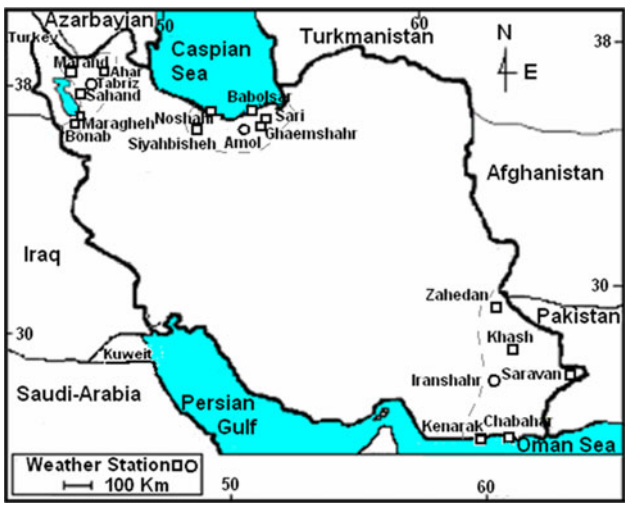


Fig. 1 Study area and location of stations in Iran (*open circles* and *open squares* show the location of targets and their closest weather stations, respectively)

where P and T are the average annual precipitation (mm) and temperature (°C), respectively. The values of aridity index of stations are shown in Table 1.

In this research, six meteorological variables namely, minimum temperature (T_{min}), maximum temperature (T_{max}), mean temperature (T_{mean}), relative humidity (RH), wind speed (WS) and precipitation (P) at monthly time scale were considered to evaluate the performance of 11 different methods in estimating missing values of these variables.

3 Materials and methods

3.1 Simple arithmetic averaging (AA)

This is the simplest method which is commonly used to fill the missing meteorological data in meteorology and climatology. Missing data are obtained by arithmetically averaging data of the closest weather stations around a station as follows:

$$V_0 = \frac{\sum_{i=1}^{n} V_i}{n} \tag{2}$$

where V_0 is the estimated value of the missing data, V_i is the value of the same parameter at ith nearest weather station and n is the number of the nearest stations, which their information were used for estimation of the missing value.

3.2 Inverse distance interpolation (ID)

The inverse distance method is used to estimate missing data because of its simplicity (Hubbard 1994).

$$V_0 = \frac{\sum_{i=1}^{n} \left(V_i/d_i\right)}{\sum_{i=1}^{n} \left(1/d_i\right)} \tag{3}$$

where d_i is the distance between the station having the missing data and the ith nearest weather station. The other parameters introduced before.

3.3 Normal ratio method (NR)

The normal ratio method which first proposed by Paulhus and Kohler (1952), and later modified by Young (1992) is a common method for estimation of rainfall missing data. The estimated data are considered as a combination of variables with different weights i.e.

$$V_0 = \frac{\sum_{i=1}^{n} W_i V_i}{\sum_{i=1}^{n} W_i} \tag{4}$$

where w_i is the weight of the ith nearest weather station and can be estimated as:

Table 1 Characteristic of selected weather stations of Iran

Province	Station	Zone code	Latitude (N)	Longitude (E)	Elevation (m)	Length of data (years)	Index of aridity	Climate type[a]
East Azarbayjan	Ahar	40704	38°26′	47°04′	1390.5	2001–2005	12.72	Semi-dry
	Bonb	99239	37°20′	46°04′	1290.0	2001–2005	11.16	Semi-dry
	Tabriz	40706	38°05′	46°17′	1361.0	2001–2005	10.66	Semi-dry
	Sahand	40707	37°56′	46°07′	1641.0	2001–2005	9.19	Dry
	Marand	99200	38°28′	45°46′	1550.0	2001–2005	15.88	Semi-dry
	Maragheh	40713	37°24′	46°16′	1477.7	2001–2005	11.60	Semi-dry
Mazandaran	Siyahbisheh	40735	36°15′	51°18′	1855.4	2001–2005	26.46	Semi-humid
	Noshahr	40734	36°39′	51°30′	−20.9	2001–2005	51.24	Extra humid
	Amol	99309	36°28′	52°23′	23.7	2001–2005	25.09	Semi-humid
	Ghaemshahr	40737	36°27′	52°46′	14.7	2001–2005	29.25	Humid
	Sari	40759	36°33′	53°00′	23.0	2001–2005	28.66	Humid
	Babolsar	40736	36°43′	52°39′	−21.0	2001–2005	37.11	Extra humid
Zahedan	Kenarak	40897	25°26′	60°22′	12.0	2001–2005	1.36	Dry
	Chabahar	40898	25°17′	60°37′	8.0	2001–2005	1.31	Dry
	Iranshahr	40879	27°12′	60°42′	591.1	2001–2005	1.52	Dry
	Saravan	40878	27°20′	62°20′	1195.0	2001–2005	2.08	Dry
	Khash	40870	28°13′	61°12′	1394.0	2001–2005	3.01	Dry
	Zahedan	40856	29°28′	60°53′	1370.0	2001–2005	1.58	Dry

[a] Climate types have been determined using De Martonne formula. Bold stations denote the target station in each climate

$$W_i = \left[r_i^2 \left(\frac{n_i - 2}{1 - r_i^2} \right) \right] \quad (5)$$

where r_i is the correlation coefficient between the target station and the ith surrounding station, n_i is the number of points used to derive the correlation coefficient.

3.4 Single best estimator (SIB)

The SIB is a simple method and analogous to using the closest neighboring station as an estimate for a target station. Target station conditions are estimated using data from the neighboring station that has the highest positive correlation with the target station.

3.5 Multiple regression analysis (REG)

The REG using the least absolute deviation criteria (MLAD) is a robust version of a general linear least squares estimation. The method of least squares is an effective method when the errors are normally distributed and independent. However, for precipitation data especially, the assumption of normality over the wide range of situations can lead to poor estimations (Eischeid et al. 1995). The main advantage of least absolute deviations is its resistance to outliers and to overemphasis of large tailed distributions (Barrodale and Roberts 1973). MLAD estimates the unknown parameters in a stochastic model so as to minimize the sum of absolute deviations of neighboring stations observations from the values predicted by the model. Kemp et al. (1983), Young (1992) and Eischeid et al. (1995) highlighted many advantages of the REG in the data interpolation and estimation of missing data. Missing data (V_0) were estimated as

$$V_0 = a_0 + \sum_{i=1}^{n} (a_i V_i) \quad (6)$$

where $a_0, a_1, \ldots, a_n$ are regression coefficients.

3.6 UK traditional method (UK)

The method traditionally used by the UK Meteorological Office to estimate missing temperature and sunshine data was based on comparisons with a single neighboring station. For temperature, a constant difference between stations was assumed. Thus, if the March temperature at Station A was 0.2°C above that at Station B averaged over a period of overlapping records, then 0.2°C was added to the recorded value at Station B to give the corresponding values at Station A. For sunshine, a constant ratio between two neighboring stations was assumed. Thus, if the January sunshine at Station A has been 1% less than at Station B during a period of overlapping records, then 1% was subtracted from the values at Station B to provide estimated values at Station A. In our study, a constant difference between stations is used for T_{min}, T_{max}, T_{mean}, RH, WS and

P. The value of Station B is an arithmetic average value resulting from the five nearest stations.

3.7 Closest station method (CSM)

In this method, the closest station was identified in the first stage. Then the missing data were estimated from the information of the closest station. In the second stage, estimated data were adjusted by the ratio of the long-term means for that month.

3.8 Cubic spline method (CSP)

The cubic spline defined over the interval [0, a] of the index variable x, is constructed in the following way. The interval [0, a] is subdivided as $0 = x_0 < x_1 < x_2 < \cdots < x_N = a$ and the corresponding set of data values $\{y_i\}$ is given. The problem is to seek a function *S(x)* which has the following properties:

1. *S(x)* along with its first and second derivatives is continuous.
2. *S(x)* is cubic within each sub-interval.
3. $S(x_i) = y_i, \quad i = 1, 2, \ldots, N.$

The spline can be defined over each sub-interval $[x_{i-1}, x_i]$ as:

$$S''(x) = M_{i-1}\frac{x_i - x}{h_i} + M_i\frac{x - x_i - 1}{h_i}, \quad x_{i-1} \le x \le x_i \tag{7}$$

in which $S''(x)$ is the second derivative of $S(x); h_i = x_i - x_{i-1}$; and M_i is known as the moment associated with the node x_i and equals the second derivative of *S(x)* at node x_i. By integrating Eq. 7 twice and evaluating the integration constants with the (boundary) conditions $S(x_{i-1}) = y_{i-1}$ and $S(x_i) = y_i$, the spline can be obtained in terms of the unknown moments M_{i-1} and M_i:

$$S(x) = M_{i-1}\frac{(x_i - x)^3}{6h_i} + M_i\frac{(x - x_{i-1})^3}{6h_i} + \frac{x_i - x}{h_i}y_{i-1} - M_{i-1}h_i^2\frac{x_i - x}{h_i} + \frac{x - x_{i-1}}{h_i}y_i - M_i h_i^2\frac{x - x_{i-1}}{6h_i} \tag{8}$$

The continuity of the first derivative of *S(x)* gives

$$\frac{h_i}{6}M_{i-1}\frac{h_i + h_{i+1}}{3}M_i + \frac{h_{i+1}}{6}M_{i+1} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \tag{9}$$

and by defining

$$h_i = \frac{h_{i+1}}{h_i + h_{i+1}}, \quad \rho_i = 1 - \lambda_i,$$
$$d_i = \frac{6}{h_i + h_{i+1}}\left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}\right)$$

Eq. 9 becomes

$$\rho_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = d_i \quad \text{for} \quad i = 1, 2, \ldots, N-1 \tag{10}$$

Since the information at the end points is missing, natural cubic spline functions such as $M_0 = M_N = 0$ are used. Therefore, Eq. 10 can be expressed in matrix form as

$$\begin{bmatrix} 2 & \lambda_1 & 0 & 0 & \ldots & 0 & 0 \\ \rho_2 & 2 & 0 & 0 & \ldots & 0 & 0 \\ 0 & \rho_3 & 2 & \lambda_3 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 2 & \lambda_{N-2} \\ 0 & 0 & 0 & 0 & 0 & \rho_{N-1} & 2 \end{bmatrix} \times \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ \vdots \\ M_{N-2} \\ M_{N-1} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{bmatrix} \tag{11}$$

By inverting the foregoing matrix, $M_i (i = 1, 2, \ldots, N-1)$ can be obtained and the cubic spline functions can be determined (Inchida and Yoshimoto 1981).

3.9 Artificial neural network (ANN)

ANNs are parallel information processing systems consisting of a set of neurons (nodes) arranged in layers and when weighted inputs are used, these nodes provide suitable conversion functions. Any layer consists of pre-designated neurons and each neural network includes one or more of these interconnected layers. Figure 2 represents a three layered structure that consists of (i) Input layer, (ii) Hidden layer, and (iii) Output layer. Further information on ANNs can be found in different textbooks, e.g. Haykin (1999).

The type of ANN used in this study is a multi-layer feed-forward perceptron (MLP) trained with the use of back

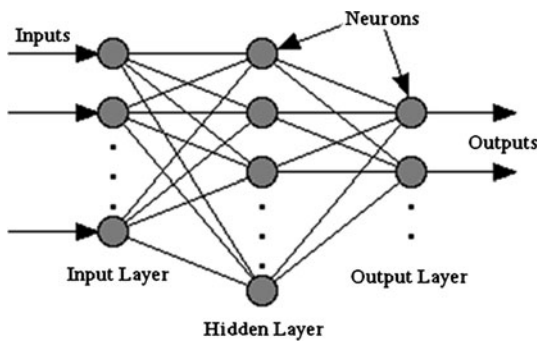


Fig. 2 Multi layer perceptron (MLP) neural network

propagation learning algorithm. The operation process of these networks is so that the input layer accepts the data and the intermediate layer processes them and finally the output layer displays the resultant outputs of the model. During the modelling stage, coefficients related to present errors in nodes are corrected through comparing the model outputs with recorded input data.

Connection weights are first initialised randomly by assigning a small positive or negative random value through the following procedure:

1. Input–output patterns are selected randomly using the training data presented to ANN.
2. Actual network outputs are calculated for the current input after the application to the activation function.
3. Performance measure is selected, e.g. MSE and the values are calculated.
4. Connection weights are adjusted to minimise the MSE.
5. Steps (2)–(5) are repeated for each pair of input–output vector in the training datasets, until no significant change in the MSE is detected for the system.

The final connection weights are kept fixed at the completion of training and new input patterns are presented to the network to produce the corresponding output consistent with the internal representation of the input/output mapping.

3.10 Adaptive nero-fuzzy inference system (ANFIS)

An adaptive network is utilised to embed the Sugeno fuzzy model into its framework to facilitate the learning of the Sugeno fuzzy model. The Sugeno fuzzy model uses a more systematic method to automatically generate the fuzzy rules, based on the input–output data, the membership functions' type and numbers and the optimization method. This can also compute gradient vectors systematically. This network architecture is called Adaptive-Network-based Fuzzy Inference System or Adaptive Neuro-Fuzzy Inference System (ANFIS). This is a universal method first introduced by Jang (1993) capable of approximating any real continuous function on a compact set to any degree of accuracy. It identifies a set of parameters through a hybrid learning rule combining the back-propagation gradient descent error and a least-squares method. It can be used as a basis for constructing a set of fuzzy "If–Then" rules with appropriate membership functions in order to generate the preliminary stipulated input–output pairs. Figure 3 represents a typical ANFIS architecture, and outline as follows:

According to Fig. 3, ANFIS model structure consists of five layers:

Layer 1: Every node in this layer is an adaptive node with a node function that may be a generalised bell or a Gaussian membership function.

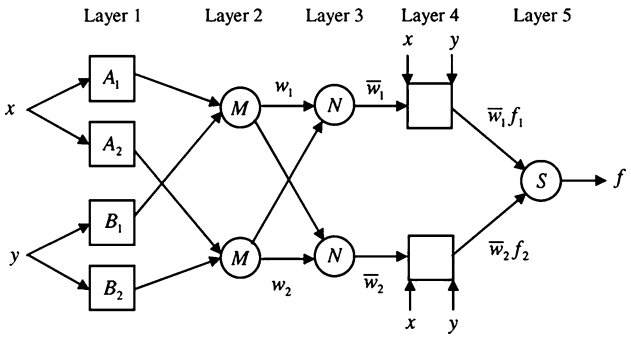


Fig. 3 A typical ANFIS architecture

Layer 2: Every node in this layer is a fixed node labelled, M, representing the firing strength of each rule (w_i), and is calculated by the fuzzy AND connective of the 'product' of the incoming signals.

Layer 3: Every node in this layer is a fixed node labelled, N, representing the normalised firing strength of each rule ($\bar{w}_i$). The ith node calculates the ratio of the ith rule's firing strength to the sum of two rules' firing strengths.

Layer 4: Every node in this layer is an adaptive node with a node function (f_i) indicating the contribution of ith rule toward the overall output.

Layer 5: The single node in this layer is a fixed node labelled, S, indicating the overall output as the summation of all incoming signals (Moghaddamnia et al. 2009).

The above comprise three different types of components, as follows (Lughofer and Klement 2003):

1. *Premise parameters* as nonlinear parameters that appear in the input membership functions.
2. *Consequent parameters* as linear parameters that appear in the rules consequents (output weights).
3. *Rule structure* that needs to be optimised to achieve a better linguistic interpretability.

3.11 Genetic programming (GP)

GP is truly a "bottom up" process, for not making any assumption on the structure of the relationship between independent and dependent variables but it identifies an appropriate relationship for any given time series. The construction of the relationship is made possible by two components and efficient emulation of evolutionary processes become possible only when these components work hand-in-hand. These components are: (i) a set of components of functions and their parameters (referred to as the terminal set), which emulates the role of proteins or chromosomes in biological systems; and (ii) a parse tree,

which is a functional set of basic operators and those selected in this study are:

$$\{+, -, \times, \div, x, x^2, x^3, \sqrt[2]{x}\}$$

The relationship between independent and dependent variables are often referred to as the "model," the "program," or the "solution." Whatever, the terminology, the identified relationship in a particular GP modelling is continually evolving and never fixed. The evolution starts from an initially selected random population of models, where the fitness value of each model is evaluated using the values of the independent and dependent variables. As the population evolves from one generation to another, new models replace the old ones by having demonstrably better performance.

There are various selection methods and the method used in this study is referred to as the gene expression programming (GEP) based on evolving computer programs of different sizes and shapes encoded in linear chromosomes of fixed lengths (Ferreira 2001a, b). The chromosomes are composed of multiple genes, each gene encoding a smaller subprogram. Furthermore, the structural and functional organisation of the linear chromosomes allows the unconstrained operation of important genetic operators, such as mutation, transposition and recombination. It has been reported that GEP is 100–10,000 times more efficient than GP systems (Ferreira 2001a, b) for a number of reasons, including: (i) the chromosomes are simple entities: linear, compact, relatively small, easy to manipulate genetically (replicate, mutate, recombine, etc.); (ii) the parse trees or expression trees are exclusively the expression of their respective chromosomes; they are entities upon which selection acts, and according to fitness, they are selected to reproduce with modification.

Applying operators like crossover and mutation to the winners, "children" or "offspring" are produced, in which crossovers are responsible for maintaining identical features from one generation to another but mutation causes a random change in the parse tree, although data mutation is also possible. This completes the operations at the initial generation and the process is repeated until termination. There are now various software applications for implementing GP models and the GeneXpro Tools (Ferreira 2001a, b) was used in this study.

4 Performance criteria

In order to compare the accuracy of the candidate methods for reconstructing missing data and selecting the most appropriate one, three statistical measures were used as follows:

(i) Mean absolute errors (MAE):

$$MAE = \frac{\sum_{i=1}^{n} |V_{obs_i} - V_{est_i}|}{n} \tag{12}$$

(ii) Coefficient of efficiency (CE):

$$CE = 1 - \frac{\sum_{i=1}^{n} (V_{obs_i} - V_{est_i})^2}{\sum_{i=1}^{n} (V_{obs_i} - V_{ave})^2} \tag{13}$$

(iii) Skill score (SS):

$$SS(V_{est}, V_{ave}, V_{obs}) = 1 - \frac{MSE(V_{est}, V_{obs})}{MSE(V_{ave}, V_{obs})} \tag{14}$$

$$MSE(V_{est}, V_{obs}) = \frac{\sum_{i=1}^{n} (V_{est_i} - V_{obs_i})^2}{n},$$
$$MSE(V_{ave}, V_{obs}) = \frac{\sum_{i=1}^{n} (V_{ave} - V_{obs_i})^2}{n}$$

where V_{obs_i}, V_{est_i} and V_{ave} are the observed, estimated and average of observed data, respectively and n is the number of missing data points which are estimated.

MAE indicates a measure of how far the estimate can be in error, ignoring sign. The method which yields the lowest value for MAE indicated as the best method for the study area. The range of MAE is from 0 to $+\infty$.

CE can range from $-\infty$ to 1. An efficiency of 1 (CE = 1) corresponds to a perfect match of estimated data to the observed. An efficiency of 0 (CE = 0) indicates that the model predictions are as accurate as the mean of the observed data, whereas an efficiency less than zero (CE < 0) occurs when the observed mean is a better predictor than the model. The closer the model efficiency is to 1, the more accurate the model is.

The skill score (SS) is positive (negative) when the accuracy of the forecasts is greater (less) than the accuracy of the reference forecasts. Moreover, SS = 1 when $MSE(V_{est}, V_{obs}) = 0$ (perfect forecasts) and SS = 0 when $MSE(V_{est}, V_{obs}) = MSE(V_{ave}, V_{obs})$. Climatology is taken to be the standard of reference. Thus, the reference forecasts are assumed to be based solely on a relevant set of observations of the variable of interest. Several alternative definitions of the climatological reference forecasts are possible, depending on the particular set of observations employed and the way in which the observations are used to create the forecasts. First, the climatological forecasts could be based on observations from some historical period (external climatological forecasts) or they could be based the sample of observations from the experimental period (internal climatological forecasts). Second, the reference forecast could consist of a single constant forecast applicable to all forecasting occasions or they could consist of different forecasts for different occasions (Murphy 1988). Generally, there are four types of climatological reference

Table 2 Performance criteria values for different methods of estimating missing data in three distinct climates of Iran

Climate	Method	T_{min}			T_{max}			T_{mean}			WS			RH			P		
		MAE	CE	SS	MAE	CE	SS	MAE	CE	SS	MAE	CE	SS	MAE	CE	SS	MAE	CE	SS
semi-dry and dry	AA	0.66	0.99	0.99	1.09	0.99	0.99	0.40	**1.00**	**1.00**	1.17	0.21	0.21	2.40	0.92	0.92	5.78	0.90	0.90
	ID	0.75	0.99	0.99	2.17	0.96	0.96	0.87	0.99	0.99	2.85	-4.06	-4.06	2.39	0.87	0.87	8.12	0.82	0.82
	NR	0.45	0.99	0.99	1.16	0.99	0.99	0.27	**1.00**	**1.00**	0.74	0.52	0.52	2.55	0.91	0.91	6.45	0.88	0.88
	SIB	0.54	0.99	0.99	2.23	0.96	0.96	0.92	0.99	0.99	3.11	-5.00	-5.00	3.92	0.86	0.86	8.56	0.80	0.80
	REG	0.34	**1.00**	**1.00**	0.29	**1.00**	**1.00**	0.50	**1.00**	**1.00**	0.46	0.84	0.84	3.57	0.90	0.90	6.94	0.86	0.86
	UK	0.81	0.99	0.99	0.33	**1.00**	**1.00**	0.58	0.99	0.99	1.27	-0.31	-0.31	2.60	0.86	0.86	8.07	0.90	0.90
	CSM	0.75	0.99	0.99	1.02	0.99	0.99	0.85	0.99	0.99	0.60	0.67	0.67	2.65	0.85	0.85	6.62	0.84	0.84
	CSP	12.20	-1.90	-1.90	18.38	-2.96	-2.96	15.00	-2.30	-2.30	2.54	-3.60	-3.60	28.69	-6.36	-6.36	72.31	-6.31	-6.31
	ANN	**0.31**	**1.00**	**1.00**	**0.23**	**1.00**	**1.00**	**0.22**	**1.00**	**1.00**	**0.29**	**0.92**	**0.92**	2.39	**0.95**	**0.95**	**3.48**	**0.97**	**0.97**
	ANFIS	0.43	**1.00**	**1.00**	0.47	**1.00**	**1.00**	0.93	0.98	0.98	1.36	-1.39	-1.39	1.97	**0.95**	**0.95**	5.78	0.86	0.86
	GP	0.34	**1.00**	**1.00**	0.31	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.31	0.91	0.91	**1.95**	**0.95**	**0.95**	6.15	0.92	0.92
semi-humid, humid and extra humid	AA	1.19	0.97	0.97	1.43	0.95	0.95	2.26	0.90	0.90	0.64	0.04	0.04	1.43	0.59	0.59	19.94	0.72	0.72
	ID	0.24	**1.00**	**1.00**	0.48	0.99	0.99	1.18	0.97	0.97	0.63	0.17	0.17	2.34	0.09	0.09	28.74	0.19	0.19
	NR	0.24	**1.00**	**1.00**	0.27	**1.00**	**1.00**	0.95	0.98	0.98	0.67	0.07	0.07	3.33	-0.72	-0.72	32.62	-0.10	-0.10
	SIB	0.37	**1.00**	**1.00**	0.32	**1.00**	**1.00**	0.67	0.99	0.99	0.70	-0.07	-0.07	5.25	-3.14	-3.14	62.79	-3.19	-3.19
	REG	0.15	**1.00**	**1.00**	0.20	**1.00**	**1.00**	0.14	**1.00**	**1.00**	0.61	0.18	0.18	0.61	0.90	0.90	14.51	0.88	0.88
	UK	0.38	**1.00**	**1.00**	0.50	0.99	0.99	0.25	**1.00**	**1.00**	0.96	-0.88	-0.88	1.93	0.18	0.18	66.44	-0.39	-0.39
	CSM	0.87	0.98	0.98	0.51	0.99	0.99	0.30	**1.00**	**1.00**	0.90	-0.70	-0.70	1.85	0.24	0.24	40.24	-2.62	-2.62
	CSP	12.65	-3.52	-3.52	10.29	-2.00	-2.00	11.07	-2.43	-2.43	6.13	-64.17	-64.17	10.47	-19.07	-19.07	64.17	-1.79	-1.79
	ANN	**0.11**	**1.00**	**1.00**	0.20	**1.00**	**1.00**	**0.13**	**1.00**	**1.00**	**0.59**	0.11	0.11	0.56	**0.94**	**0.94**	**8.54**	**0.95**	**0.95**
	ANFIS	0.29	**1.00**	**1.00**	0.65	0.99	0.99	0.21	**1.00**	**1.00**	0.61	0.05	0.05	0.91	0.81	0.81	10.08	0.89	0.89
	GP	0.12	**1.00**	**1.00**	**0.18**	**1.00**	**1.00**	0.20	**1.00**	**1.00**	**0.59**	**0.25**	**0.25**	**0.52**	0.93	0.93	11.95	0.93	0.93
dry	AA	3.80	0.73	0.73	5.17	0.54	0.54	4.06	0.70	0.70	0.69	-0.17	-0.17	16.93	-5.62	-5.62	3.67	-1.85	-1.85
	ID	4.68	0.65	0.65	5.39	0.55	0.55	4.11	0.73	0.73	0.70	-0.17	-0.17	11.15	-1.83	-1.83	3.90	-1.71	-1.71
	NR	4.14	0.71	0.71	5.96	0.50	0.50	4.57	0.71	0.71	0.66	-0.13	-0.13	1.98	0.89	0.89	3.95	-1.90	-1.90
	SIB	5.55	0.55	0.55	6.63	0.39	0.39	4.72	0.69	0.69	0.96	-1.47	-1.47	1.58	0.93	0.93	5.35	-3.65	-3.65
	REG	0.65	**0.99**	**0.99**	0.32	**1.00**	**1.00**	0.40	**1.00**	**1.00**	0.38	0.61	0.61	1.19	0.95	0.95	4.03	-0.97	-0.97
	UK	1.17	0.98	0.98	0.40	**1.00**	**1.00**	0.40	**1.00**	**1.00**	0.88	-0.86	-0.86	3.75	0.58	0.58	6.45	-1.19	-1.19
	CSM	3.20	0.82	0.82	1.56	0.96	0.96	1.37	0.97	0.97	0.79	-0.38	-0.38	4.54	0.405	0.405	3.77	-3.62	-3.62
	CSP	13.97	-2.87	-2.87	13.84	-2.75	-2.75	14.38	-2.99	-2.99	0.54	0.05	0.05	34.91	-29.97	-29.97	21.30	-36.73	-36.73
	ANN	**0.49**	**0.99**	**0.99**	0.31	**1.00**	**1.00**	0.41	**1.00**	**1.00**	0.40	0.59	0.59	1.13	**0.97**	**0.97**	**2.32**	**0.33**	**0.33**
	ANFIS	0.52	**0.99**	**0.99**	0.69	0.99	0.99	**0.39**	**1.00**	**1.00**	1.59	-6.73	-6.73	8.95	-2.84	-2.84	2.57	0.20	0.20
	GP	0.58	**0.99**	**0.99**	**0.30**	**1.00**	**1.00**	0.40	**1.00**	**1.00**	**0.33**	**0.70**	**0.70**	**1.08**	0.96	0.96	3.10	-0.18	-0.18

Note: some explanation about the abbreviations of the methods names in column two is as follows:
AA: Simple Arithmetic Averaging; *ID*: Inverse Distance Interpolation; *NR*: Normal Ratio Method; *SIB*: Single Best Estimator; *REG*: Multiple Regression Analysis; *UK*: UK Traditional Method; *CSM*: Closest Station Method; *CSP*: Cubic Spline Method; *ANN*: Artificial Neural Network; *ANFIS*: Adaptive Neuro-Fuzzy Inference System; and *GP*: Genetic Programming.
Highlighted cells indicate appropriate method among all eleven candidate methods based on a certain criterion for a given meteorological variable (see text for more information).

forecasts and Eq. 14 is for single-valued internal reference forecasts which are suitable to be considered in this study. For more information about climatological reference forecasts, see Murphy (1974, 1988).

5 Results

The three stations namely Tabriz, Amol and Iranshahr located in three distinct climates of Iran (Fig. 1) were considered as target stations. To evaluate the efficiency of the methods, the whole monthly data of a year (2002 here) were re-estimated by different 11 candidate methods expressed in the previous section. This was repeated for all the six meteorological parameters. The year 2002 was chosen because the record was complete for that year. In the case of three methods, ANN, ANFIS and GP, the information of the year 2002 were not used in the training or calibration phases. Three goodness-of-fit methods including MAE, CE, and SS are considered to select the appropriate method. For the six meteorological variables, results of the methods are presented in Table 2 for the three different climates. In Table 2, we highlighted the cells having the least MAE and highest CE and SS. Such highlighted cells indicate appropriate method among the 11 candidate methods for a given meteorological parameter and climate condition. This helps to identify the most suitable method for each meteorological parameter and climatic region. For example, according to Table 2, it can be seen that ANN has the least MAE and the highest CE and SS values for minimum temperature (T_{min}) among all other methods for all three climatic regions which implies that the ANN is the most appropriate method for estimating T_{min} variable for different climate conditions. Figure 4 represents the selected suitable methods evaluation of missing values vis-a-vis their actual observations for the six variables of three target stations (Tabriz, Amol and

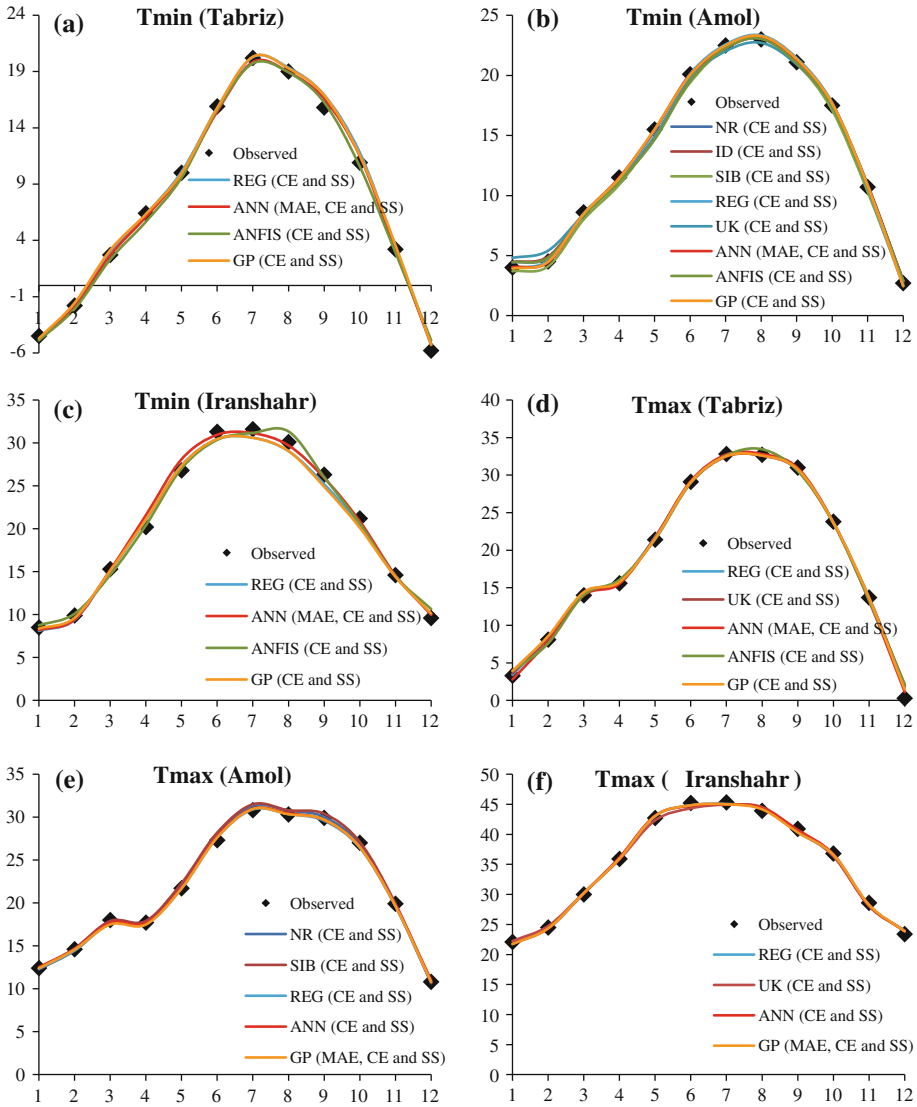
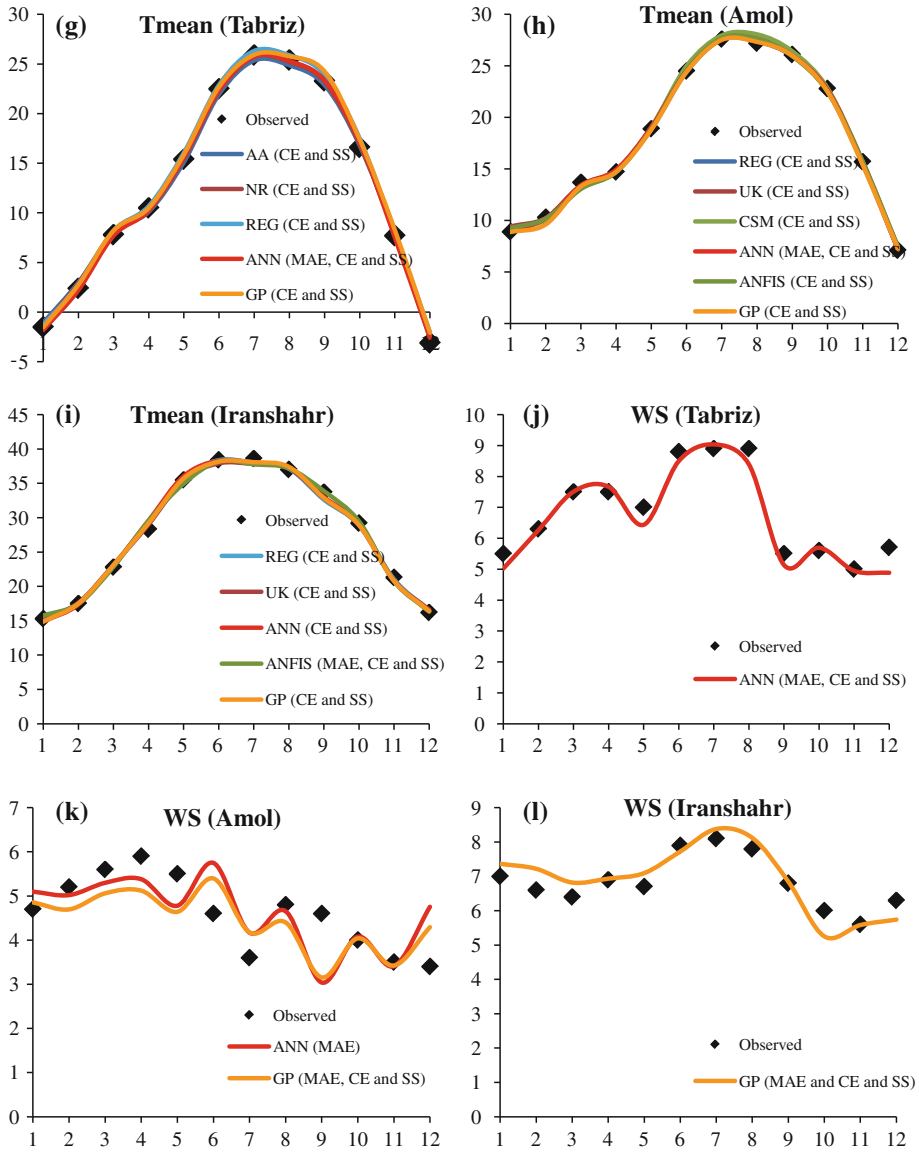


Fig. 4 Evaluation of missing values by the suitable methods. *Note*: units of T_{min}, T_{max} and T_{mean} is in °C; WS is in m/s; RH is dimensionless (%) and P is in mm. Numbers in *horizontal axes* denote months (i.e. 1: January, 2: February, …). Note that most of the curves coincide each other in the figures

Fig. 4 continued



Iranshahr). As it can be seen from Fig. 4, the selected methods show high ability and relatively similar performance in estimating the related variable.

Comparison of methods:

In this section, the selected methods among all 11 candidate methods are introduced according to Table 2 and based on performance criteria indicated within parenthesis. This was accomplished for all six meteorological variables as well as three distinct climates of Iran.

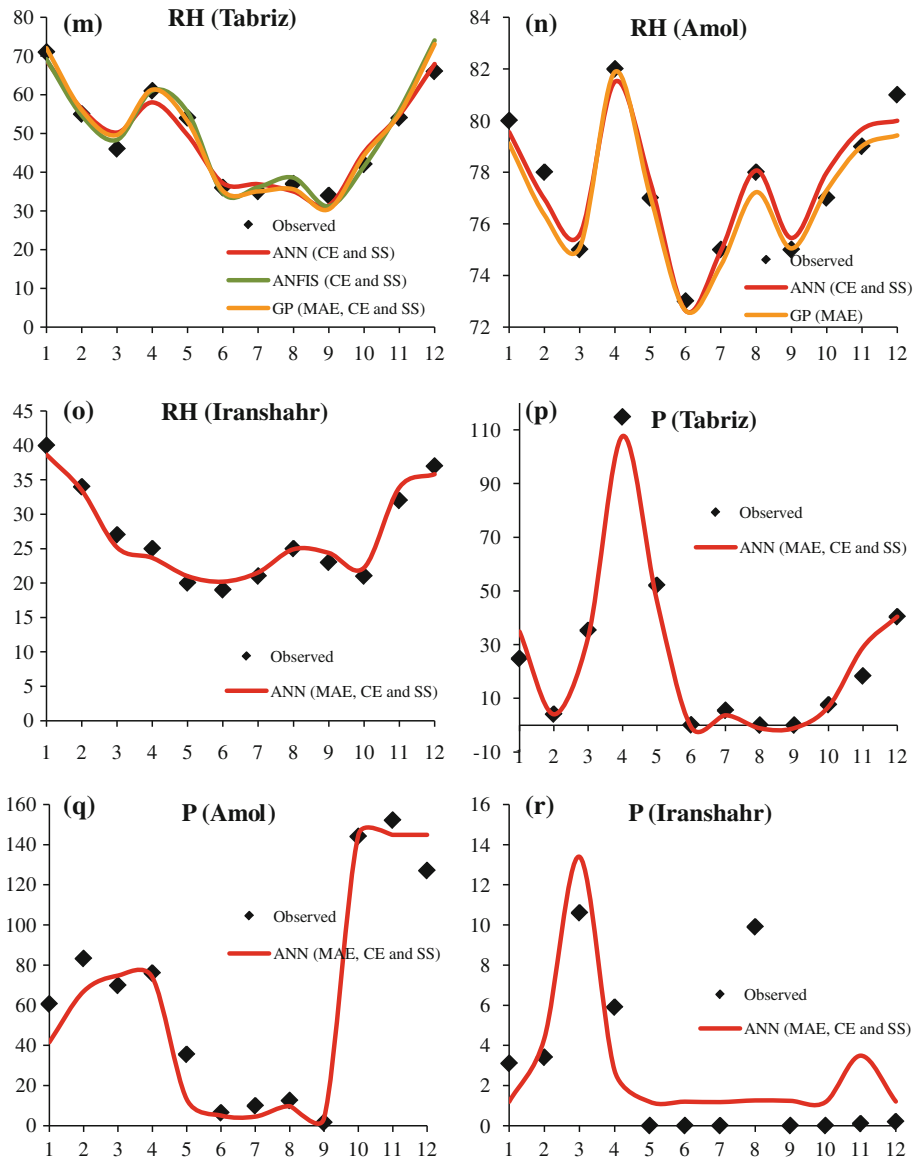
For semi-dry and dry climate located in the north west of Iran, the appropriate methods are found to be as follows:

T_{min}: REG (CE and SS), ANN (MAE, CE and SS), ANFIS (CE and SS) and GP (CE and SS);
T_{max}: REG (CE and SS), UK (CE and SS), ANN (MAE, CE and SS), ANFIS (CE and SS) and GP (CE and SS);
T_{mean}: AA (CE and SS), NR (CE and SS), REG (CE and SS), ANN (MAE, CE and SS) and GP (CE and SS);
WS: ANN (MAE, CE and SS);
RH: ANN (CE and SS), ANFIS (CE and SS) and GP (MAE, CE and SS);
P: ANN (MAE, CE and SS).

For semi-humid, humid and extra humid climate located in the north of Iran, the appropriate methods are found to be as follows:

T_{min}: ID (CE and SS), NR (CE and SS), SIB (CE and SS), REG (CE and SS), UK (CE and SS), ANN (MAE, CE and SS), ANFIS (CE and SS) and GP (CE and SS);
T_{max}: NR (CE and SS), SIB (CE and SS), REG (CE and SS), ANN (CE and SS) and GP (MAE, CE and SS);

Fig. 4 continued



T_{mean}: REG (CE and SS), UK (CE and SS), CSM (CE and SS), ANN (MAE, CE and SS), ANFIS (CE and SS) and GP (CE and SS);
WS: ANN (MAE) and GP (MAE, CE and SS);
RH: ANN (CE and SS) and GP (MAE);
P: ANN (MAE, CE and SS).

For dry climate located in the south east of Iran, the appropriate methods are found to be as follows:

T_{min}: REG (CE and SS), ANN (MAE, CE and SS), ANFIS (CE and SS) and GP (CE and SS);
T_{max}: REG (CE and SS), UK (CE and SS), ANN (CE and SS) and GP (MAE, CE and SS);
T_{mean}: REG (CE and SS), UK (CE and SS), ANN (CE and SS), ANFIS (MAE, CE and SS) and GP (CE and SS);
WS: GP (MAE, CE and SS);
RH: ANN (MAE, CE and SS);
P: ANN (MAE, CE and SS).

6 Discussion

Based on the results obtained, it may be inferred that the ANN, GP and ANFIS (or REG) methods are generally suitable for estimating missing data of semi-dry and dry climate of Iran. For semi-humid, humid and extra humid climate and also for dry climate, however, the ANN, GP and REG methods seem to be more suitable methods. Taking these collectively, the ANN and GP methods may be suggested as appropriate methods for the three climates of Iran, and possibly for other climates of the world, although caution needs to be exercised in making such a generalization. The suitability of the ANN and GP methods

for estimating missing wave height data has been proved by Ustoorikar and Deo (2008).

The results also indicate that the REG method is an appropriate method among the eight traditional methods for all variables and different climate conditions. This is in accord with the findings of Eischeid et al (1995) and Xia et al. (1999). Furthermore, Shih and Cheng (1989) stated that the regression technique and the regional average can be successfully applied to generate missing monthly solar radiation data at Puerto Rico. Unfortunately, they did not use recently developed methods such as ANN, ANFIS and GP. They found that the regression technique and the averaging of historical data have been satisfactorily used to interpolate missing RH values. The CSP method is the worst in estimating all variables in different climates except for wind speed variable in semi-dry, dry climate and dry climate of Iran. However, Mizumura (1985) and Price et al. (2000) showed that CSP is an efficient method for missing data infilling. Moreover, the traditional methods show relatively good ability just in estimating T_{min}, T_{max} and T_{mean} rather than the WS, RH and P variables in all climates conditions.

Although in this study the appropriate methods for estimating missing values of meteorological parameters were selected for Iran, in our opinion the results would be applicable for other arid and semi-arid countries. This is due to the fact that all arid and semi-arid regions have the same climate conditions.

7 Conclusion

In this study, the ability of 11 different methods in estimating various missing climatological data in monthly time scale was evaluated in three different climates of Iran. The performances of these methods for different variables and climates were compared using mean absolute error (MAE), coefficient of efficiency (CE) and skill score (SS). The results generally suggest that, among the 11 different missing data estimation methods, the ANN method is the most appropriate model for estimating missing values of different climate conditions of Iran, followed by the GP method. The high capability of the GP and ANN methods has been reported by many researchers such as Ustoorikar and Deo (2008) and Teegavarapu and Chandramouli (2005). As Dastorani et al. (2009) suggested, the artificial intelligence techniques are more powerful than the traditional methods in missing data estimation. It can be concluded that selection of appropriate method for estimating missing climatological data is very important, because the difference of the evaluation criteria values estimated by the best (ANN and GP) and worst (CSP) methods is relatively too high.

In this research, three distinct climates of Iran were examined. Conducting similar studies is strongly recommended for other climate conditions of the world. Also, it is suggested that such a study should be carried out for identifying suitable method in missing data infilling for other hydrological variables like solar radiation, pan evaporation, reference crop evapotranspiration, runoff and sediment load and other time scales such as daily, weekly and biweekly.

References

Abebe AJ, Solomatine DP, Venneker RGW (2000) Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. Hydrol Sci J 45(3):425–436

Barrodale I, Roberts FDK (1973) An improved algorithm for discrete L1 approximation. SLAM J Num Anal 10:839–848

Bussieres N, Hogg W (1989) The objective analysis of daily rainfall by distance weighting schemes on a mesoscale grid. Atmos Ocean 27:521–541

Coulibaly P, Evora ND (2007) Comparison of neural network methods for infilling missing daily weather records. J Hydrol 341:27–41

Dastorani MT, Moghadamnia A, Piri J, Rico-Ramirez M (2009) Application of ANN and ANFIS models for reconstructing missing flow data. Environ Monit Assess. doi:10.1007/s10661-009-1012-8

Degaetano AT, Eggleston KL, Knapp WW (1995) A method to estimate missing maximum and minimum temperature observations. J Appl Meteorol 34:371–380

Eischeid JK, Baker CB, Karl TR, Diaz HF (1995) The quality control of long-term climatological data using objective data analysis. J Appl Meteorol 34:2787–2795

Ferreira C (2001a) Gene expression programming in problem solving. In: 6th Online world conference on soft computing in industrial applications (invited tutorial)

Ferreira C (2001b) Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst 13(2):87–129

Haykin S (1999) Neural networks: a comprehensive foundation, 2nd edn. Prentice Hall, Upper Saddle River

Hubbard KG (1994) Spatial variability of daily weather variables in the high plains of the USA. Agric For Meteorol 68:29–41

Huth R, Nemesova I (1995) Estimation of missing daily temperature: can a weather categorization improve its accuracy. J Appl Meteorol 34:1901–1916

Inchida K, Yoshimoto F (1981) Spline functions and its applications. Kyoiku Shuppan, Tokyo, Japan

Jang JR (1993) Anfis: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23:665–685

Kemp WP, Burnell DG, Everson DO, Thomson AJ (1983) Estimating missing daily maximum and minimum temperatures. J Clim Appl Meteorol 22:1587–1593

Kim TW, Ahn H (2009) Spatial rainfall model using a pattern classifier for estimating missing daily rainfall data. Stoch Environ Res Risk Assess 23:367–376

Lughofer E, Klement E (2003) Online adaptation of Takagi-Sugeno fuzzy inference systems. In: Proceedings of CESA'2003—IMACS multi conference, Lille, France, CD-Rom, paper S1-R-00-0175

Luo Z, Wahba G, Johnson DR (1998) Spatial-temporal analysis of temperature using smoothing spline ANOVA. J Clim 11:18–28

Mizumura K (1985) Estimation of hydraulic data by spline functions. J Hydraul Eng 9(111):1219–1225

Moghaddamnia A, Remesan R, Hassanpour Kashani M, Mohammadi M, Han D, Piri J (2009) Comparison of LLR, MLP, Elman, NNARX and ANFIS Models-with a case study in solar radiation estimation. J Atmos Sol Terr Phys 71:975–982

Murphy AH (1974) A sample skill score for probability forecasts. Mon Weather Rev 102:48–55

Murphy AH (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon Weather Rev 116:2417–2424

Paulhus JLH, Kohler MA (1952) Interpolation of missing precipitation records. Mon Weather Rev 80:129–133

Price DT, McKenney DW, Nalder IA, Hutchinson MF, Kesteven JL (2000) A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agric For Meteorol 101:81–94

Saborowski J, Stock R (1994) Regionalization of precipitation data in the Harz Mountains. Allg Forst Jagdztg 165:117–122

Shih SF, Cheng KS (1989) Generation of synthetic and missing climatic data for Puerto Rico. Water Resour Bull 25(4):829–836

Srikanthan R, Harrold TI, Sharma A, McMahon TA (2005) Comparison of two approaches for generation of daily rainfall data. Stoch Environ Res Risk Assess 19:215–226

Teegavarapu RSV, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312: 191–206

Ustoorikar K, Deo MC (2008) Filling up gaps in wave data with genetic programming. Mar Struct 21:177–195

Wallis JR, Lettenmayer DP, Wood EF (1991) A daily hydroclimatological data set for the continental United States. Water Resour Res 27:1657–1663

Willmott CJ, Robeson SM (1995) Climatologically aided interpolation (CAI) of terrestrial air temperature. Int J Climatol 15: 221–229

Xia Y, Fabian P, Stohl A, Winterhalter M (1999) Forest climatology: estimation of missing values for Bavaria, Germany. Agric For Meteorol 96:131–144

Young KC (1992) A three-way model for interpolating monthly precipitation values. Mon Weather Rev 120:2561–2569