Environment, Development and Sustainability
https://doi.org/10.1007/s10668-021-01784-4

Improving estimation of missing data in historical monthly precipitation by evolutionary methods in the semi-arid area

Mahboobeh Farzandi[1] · Hossein Sanaeinejad[1] · Hojat Rezaei-Pazhan[2] · Majid Sarmad[3]

Received: 14 June 2020 / Accepted: 21 August 2021

Abstract
Precipitation is among the main variables in weather and climate studies. The length of the statistical period plays a pivotal role in the accurate analysis of precipitation. One of the limitations of meteorological stations is having missing data. Analysis based on incomplete data leads to biased analysis. The historical monthly precipitation of the five stations in Iran is available since 1880 with missing data. The name of these synoptic stations are Mashhad, Isfahan, Tehran, Bushehr, and Jask. The data in the period of 1941–1949 have a gap that was during and following World War II (1939–1945). The present study aimed to use several classic and meta-heuristic methods to estimate these missing data. The Root Means Square Error (RMSE) criteria were used for comparison. The neighboring stations of Iran were selected as independent variable to estimate missing rainfall data. First, missing data were restored with the fitting of several new regression models for monthly precipitation (with RMSEs: 9.79, 7.89, 13.43, 6.65, and 20.96 millimeter(mm)). Then, the parameters of regression models were optimized by methods of genetic algorithm (GA) and Ant Colony (ACO). It was observed that RMSEs reduced to 2.56, 2.51, 3.49, 2.48, and 4.02 mm. Besides, Artificial Neural Network (ANN) and Support Vector Regression (SVR) methods were used to model the data. ANN and SVR could not increase the accuracy of the estimated data. The missing data were imputed using evolutionary methods (GA and ACO). As a result, the length of the statistical period of the stations reached over 125 years, and the data could be considered a valuable basis for water resources, drought analyses, evaluation trends, climate changes, and global warming.

Keyword Data assimilation · Precipitation of Iran · Missing data · Genetic algorithm · Ant colony

✉ Mahboobeh Farzandi
mhb_farzandi@yahoo.com

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

[2] Department of Civil Engineering, Faculty of Engineering, Islamic Azad University, Mashhad Branch, Mashhad, Iran

[3] Department of Statistic, Faculty of Mathematics, Ferdowsi University of Mashhad, Mashhad, Iran

Published online: 21 September 2021

1 Introduction

Missing data are among the major issues in data mining and pattern recognition. The concept of missing values is essential to the successful management of data. The knowledge in attributes and missing data is also critical in improving the decision-making process of an organization (Little & Rubin, 2002). Problems with missing data in climatic series often arise and are caused by many circumstances, mainly due to the sources of acquisition, which are usually reports, manual collection instruments, or remote sensors. Typically, these problems lead to a combination of random and chronological missing data patterns in precipitation time series (Aguilera et al., 2020). Simply ignoring missing data can lead to partial and biased results in data analysis. One of the main constraints in this regard is that meteorological stations are riddled with missing climatic data. Classical statistical analyses are mainly based on complete sample data. Analysis based on incomplete data leads to biased estimates, and bias tends to be higher with increased missing ratio (Little & Rubin, 2002). The sample size (record length) of arid and semi-arid lands should be at least 100 years (Jacob et al., 1999) for statistical analysis. Moreover, studies with limited precipitation periods cannot provide a great perspective (Belala et al., 2018).

Precipitation is an essential variable in meteorology, climatology, hydrology, and environmental sciences (Türkeş et al., 2016). This factor has a direct correlation with the regional climate. Recording of precipitation data in Iran began in 1951. These data are available on the website of Iran Meteorology Organization (IMO; http://www.irimo.ir). Accordingly, the maximum length of these records is 68 years.

Five stations in Iran have longer monthly precipitation records with missing data since 1878 (Smithsonian Institution 1927, 1934, & 1947). These stations are located in Mashhad, Tehran, Isfahan, Jask, and Bushehr. The most prominent data missing in these stations were during and following World War II (1941–1949). Consequently, the only acceptable precipitation data for analysis are available in these five stations after repair.

Several studies have been focused on the management of missing data. Some researchers have only studied classical methods. To assess suitability of the different methods for filling in missing data, Sattari et al. considered monthly precipitation data collected at six different stations. They considered various classic techniques (the arithmetic averaging method, the multiple linear regression method, and the nonlinear iterative partial least-squares algorithm) for filling in missing precipitation data. The multiple imputation method produced the most accurate results for precipitation data from five dependent stations (Sattari et al., 2017). Another study assessed the variations in daily precipitation characteristics during 1960–2014 in the source region of the Yellow River in China. Their data had missing values, and they filled the missing precipitation using the linear regression method with the stations nearby. In detail, the station with the missing data was regarded as dependent station, and its neighboring stations (without missing values) were considered as independent stations (Iqbal et al., 2018).

Coulibaly and Evora investigated six types of artificial neural network (ANN), including the multilayer perceptron (MLP) network and its variations (time-lagged feed-forward network [TLFN]), generalized radial basis function network, recurrent neural network and its variations (time-delay recurrent neural network), and counter propagation fuzzy-neural network (CFNN) using various optimization methods for infilling the total daily missing precipitation records. According to the findings, MLP, TLFN, and CFNN could provide the most accurate estimates of the missing daily precipitation data (Coulibaly & Evora, 2007).

Using precipitation data of the nearby stations is a common approach to repairing missing values. In a study, researchers reconstructed daily precipitation data series using classic models, such as generalized linear modeling (GLM). In addition, they used the rainfall data (occurrence and rate) in 10 nearby areas as dependent variable, as well as the geographic data of each station (latitude, longitude, and elevation) as independent variables (Serrano-Notivoli et al., 2017). In another research, the missing flow data were predicted by using the neighbor sites. The researchers used ANN, and Adaptive Neuro-Fuzzy Inference System (ANFIS), and conventional methods (correlation and normal ratio method). According to the results, the four methods presented acceptable predictions in some cases and the ANFIS technique showed superior ability for predicting the missing flow data, especially in arid zones. Furthermore, comparison of the results indicated that ANN was a more efficient method to predict the missing data as opposed to the conventional approaches (Dastorani et al., 2010).

Researchers have investigated meteorological drought in the northern and northwestern regions in Mexico in various climate change scenarios. In this study, a feed-forward artificial neural network approach was employed for the interpolation of the missing rainfall data (Escalante-Sandoval & Nuñez-Garcia, 2017). Another study proposed a new methodology for imputing the missing attribute values through integrating GA techniques and decision tree learning for the imputation of the missing attribute values (Patil & Bichkar, 2010).

Other methods are also available for the management of missing values. Yozgatligil et al. (2013) compared several imputation methods to complete the missing values of spatiotemporal meteorological time series. Among these methods, simple arithmetic average, normal ratio (NR), and NR-weighted correlations were considered as simple methods, whereas multilayer perceptron neural network and the multiple imputation strategy of Markov chain Monte Carlo expectation–maximization (MCMC-EM) algorithm were considered as the computationally intensive techniques. In addition, the authors proposed modification on the MCMC-EM method and concluded that using the MCMC-EM algorithm for the imputation of missing values before the statistical analysis of meteorological data could decrease uncertainty and provide robust results (Yozgatligil et al., 2013). A fixed functional set genetic algorithm method (FFSGAM) is proposed for estimating historical daily missing precipitation data of 15 rain gaging stations from the state of Kentucky, USA. This research uses genetic algorithms and a nonlinear optimization formulation to obtain optimal functional forms and coefficients. The tests of FFSGAM at two rainfall gauging stations indicated that better estimates of precipitation can be obtained compared to those from a traditional inverse distance weighting technique (Ramesh et al., 2009).

Missing data are among the top problems in data analysis and pattern recognition. Undoubtedly, the concept of missing values is essential in data management. In order to predict the missing values in the five weather stations in Iran, we could not found comparison of various techniques (e.g., classic, evolutionary, and machine learning methods) for the imputation of missing monthly rainfall data in the literatures. Therefore, we used and compared multiple regression, ANN, SVR, GA, and ACO to fill in this research gap. GA and ACo algorithms can be used to select optimally parameters in the regression patterns (Seyyednezhad Golkhatmi et al., 2012). This fact can be effective in increasing the accuracy of estimating missing data that is evaluated in this study. The present study aimed to implement and compare several classic and heuristic methods to estimate the missing data for five long-term monthly precipitation stations in Iran. Initially, several multiple regressions were made fit to each monthly station precipitation, and the optimal regression model was selected for each station. Following that, the GA and ACO were applied to improve

the accuracy of the selected regression models by optimizing their parameters. In addition, ANN and support vector regression (SVR) were used to calculate the missing monthly values. Finally, the applied methods and the selected predictors for the filling of the missing values were compared.

2 Materials and methods

This study aims to increase the accuracy of estimating the missing historical monthly rainfall data of the five stations in Iran. This research is important in two ways. 1—Using different efficient methods in increasing the accuracy of estimating missing data, 2—Collecting, reconstructing and presenting historical precipitation data of the five stations in Iran that have not been available to researchers so far. The study area and methods used are as follows.

2.1 Study site

Persia officially the Islamic Republic of Iran is a country in Western Asia in the Middle East. Iran is bordered to the northwest by Armenia and the Republic of Azerbaijan, to the north by the Caspian Sea, to the northeast by Turkmenistan, to the east by Afghanistan and Pakistan, to the south by the Persian Gulf and the Gulf of Oman, and the west by Turkey and Iraq.

Most climatic regions in Iran are arid and semi-arid. (Salehnia et al. 2017; Golkar Hamzee Yazd et al., 2019; Kazemzadeh & Malekian, 2018).

The observed data in the first synoptic station in Iran were available since 1951 (http://irimo.ir/). Long-term historical monthly precipitation data are available in five cities in Iran (Fig. 1), which have been measured and recorded by the Embassy of the United States and England since 1880 (Smithsonian Institution, 1927, 1934, & 1947). These stations are located in Mashhad, Tehran, Isfahan, Jask, and Bushehr (Fig. 1). Unfortunately, the data have missing monthly values, the most important of which were during and following World War II (1941–1949).

Due to the distance, relationship, and completeness of data since 1880, the stations in the neighboring countries were selected as the predictive variables (Fig. 1). The data of the predictive stations were used to estimate the missing monthly values in Mashhad, Tehran, Isfahan, Jask, and Bushehr. The predictive stations are located in Turkmenistan (Ashgabat, Sarakhs, and Kooshka), Iraq (Baghdad, Basra, and Diwaniya), Azerbaijan (Lenkoran), Bahrain (Bahrain), and the United Arab Emirates (Sharjah) (http://sdwebx.worldbank.org, https://climexp.knmi.nl). The features of these stations are presented in Table 1.

2.2 Statistical analysis

The framework of data analysis involved the use of several conventional and heuristic methods to estimate the missing data of the long-term monthly precipitation in the mentioned stations in Iran consisting of four stages.

At the first stage, several multiple regressions were fitted to each monthly station precipitations, and the optimal regression model was selected for each station. Afterward, GA, and ACO were applied to improve the accuracy of the selected regression models by optimizing their parameters. At the next stage, ANN and SVR were used to estimate the

Fig. 1 Location of the study area, 5 stations of Iran (bold circles) that have missing data and base stations (hollow circles)

Table 1 Names, information, positions and missing percent of selected stations

Type of variable	Stations name	Country	Long (°)	Lat (°)	Alt (m)	Duration to 2017 (year)	Missing (%)
Dependent	Mashhad(RMas)	Iran	59.63	36.27	980	1893	9.1
	Isfahan (REsf)	Iran	51.70	32.70	1590	1894	9.5
	Tehran (RTeh)	Iran	51.40	35.70	1191	1884	12.4
	Jask (RJas)	Iran	57.50	25.80	4	1893	19.5
	Bushehr (RBus)	Iran	50.80	29.00	14	1878	19.6
Independent	Ashgabat	Turkmenistan	58.33	37.97	227	1892	19.2
	Sarakhs	Turkmenistan	61.22	36.53	279	1902	29.2
	Kushkah	Turkmenistan	62.35	35.28	57	1897	25.5
	Baghdad	Iraq	44.40	33.30	34	1888	34.4
	Basreh	Iraq	47.70	30.40	2	1921	43.6
	Diwania	Iraq	45.00	32.00	20	1940	68.8
	Lenkoran	Azerbaijan	48.83	38.73	− 13	1847	32.5
	Bahrain	Bahrain	50.65	26.27	2	1902	17.7
	Sharjah	Emarat	55.50	25.30	34	1933	48.7

missing monthly values separately. Finally, the results of the previous stages were compared using the RMSE, and the optimal models were applied to determine the missing values of each station.

2.2.1 Artificial neural networks (ANN)

The ANN is derived from natural learning systems. ANN is an interconnected group of artificial neurons that uses a mathematical model for information processing based on a connectionist approach to computation. In more practical terms, neural networks are nonlinear statistical data-modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. In many applications, modeling tools have provided better results when used in hydrological time series analysis. Neural networks must be trained with a set of typical input/output pairs of data called the training set (Miang Kueh & Kuok Kuok, 2016). By attempting to map the intrinsic relationships between the data with the training process and with the neurons, it tries to provide a mapping between the input space (input layer) and the desired space (output layer). The layer (or hidden layers) processes the information received from the input layer and provides the output layer (Fig. 2). Each network trains by receiving examples. Training is a process that ultimately leads to learning.

Network training is done when the communication weights between the layers change so that the difference between the predicted and calculated values is acceptable. Learning is achieved by achieving these conditions of the process. These weights represent memory and network knowledge.

The final weight vector of a successfully trained neural network represents its knowledge about the problem. As different types of neural network deal with the issues in different ways, their ability varies depending on the nature of the problem in hand. Multilayer Perceptron networks (MLP) are a static architecture of neural networks, as well as recurrent and time-lagged recurrent neural networks, which are dynamic networks (Dastorani et al., 2010). MLP has been applied to distinct areas, performing tasks such as fitting function and pattern recognition problems, by using the supervised training with an algorithm known as "Error backpropagation." Therefore, MLP with one input layer, three hidden layers, and one output layer were used in this study. The hyperbolic tangent sigmoid is used as the activation function for the hidden nodes. The Loewenberg Marquart was selected for the training algorithm in this study.

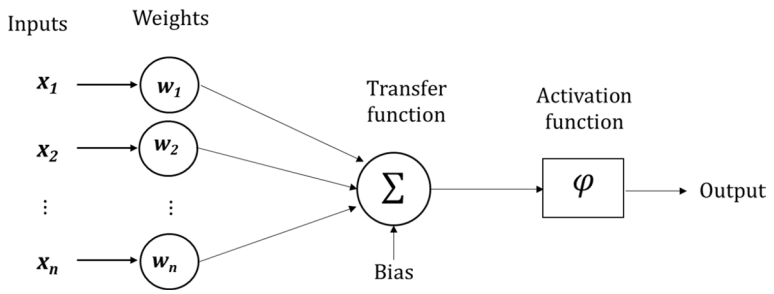

Fig. 2 A framework of a single ANN neuron

2.2.2 Support vector regression (SVR)

Support Vector Machine (SVM) is used for both classification and estimation of data fitting function in regression problems, so that the least error occurs in data classification or fitting function. This method is based on statistical learning theory, which uses the principle of structural error minimization and produces a general optimal solution. Support vector regression (SVR) is directly derived from SVM theory (Smola & Vishwanathan, 2008). Simply put, SVR is a linear regression that uses a margin instead of a line. Points near this margin are more important than farther points. Whereas linear regression considers the importance of all points equally. Both linear regression and SVR are actually data separators (Smola & Vishwanathan, 2008; Aydilek & Arslan, 2013).

SVR maintains all the main features that characterize the algorithm (maximal margin). SVR uses the same principles as the SVM for classification, with only a few minor differences. In the SVR, a Safety margin (Ɛ) is set in approximation to the SVM, which would have already requested from the problem. This algorithm is more complicated, therefore to be taken into consideration. However, the main idea is always the same: to minimize error, to individualize the hyperplane which maximizes the margin and keeping in mind that part of the error is tolerated.

Equation 1 is the Vapnik's cost function. Figure 3 depicts the situation graphically. The SVR function can be linear and nonlinear that linear type is shown in Fig. 3. All samples that fall outside the support vectors (lines with ε interval) are penalized by the cost function.

$$|y - f(x)|_{\varepsilon} = \begin{cases} 0 & \text{if } |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & \text{otherwise} \end{cases} \tag{1}$$

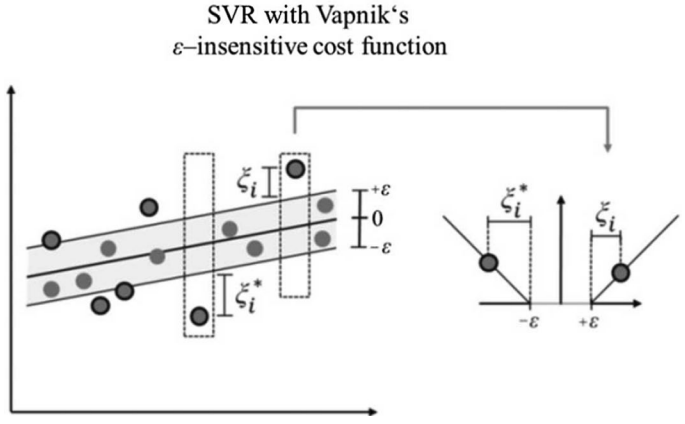

Fig. 3 In SVR, all samples outside a fixed tube with size ε (i.e., support vectors) are penalized by applying a cost function. Here, Vapnik's ε cost function is deployed, which accounts for a linear penalization

2.2.3 Genetic algorithms optimization (GA)

We use the method described by Aydilek & Arslan (2013) shortly. GA was introduced for natural selection, in which the law of survival of the fittest is applied to a population of individuals. This method is based on the biological evolution Darwin theory, and the results obtained from this study improve during the process. GA has been widely used as an effective search technique to perform searches ranging from general to specific and from simple to complex. This natural method is used for optimization. GA is implemented by generating a population and creating a new population by performing the following procedures: reproduction, crossover, and mutation. Crossover is the most significant phase in a genetic algorithm. For each pair of parents to be mated, a crossover point is chosen at random from within the genes. Mutation occurs to maintain diversity within the population and prevent premature convergence. Figure 4 shows the chromosome, gene, population and crossover and mutation operators in GA algorithm. In a GA, a population of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered; traditionally, solutions are represented in binary as strings of 0 s and 1 s, but other encodings are also possible (Salehnia et al., 2019).

The evolution usually starts from a population of randomly generated individuals and is an iterative process, with the population in each iteration called a generation.

For each member of the population, a value is assigned to represent the degree of its adaptation to the objective function. The more compatible the members are, the more likely they are to be selected and transferred to the next generation (elitism). The genetic mating function combines the genes of the chromosomes together. But it does not necessarily apply to all disciplines. Apply it with probability p. The mutation operator is performed by randomly selecting multiple chromosomes and randomly selecting one or more genes and replacing it with a reverse. The mutation is performed with a certain probability of pm. In this way, the process cycle will continue in subsequent generations. The end of the algorithm process is to achieve optimal solutions (Aydilek & Arslan 2013).

In this research, the GA is used to estimate and optimize the parameters of the regression models. The objective function is to minimize the root mean square error (RMSE). The error function is optimized by repeating the algorithm and producing better parameters in each generation.

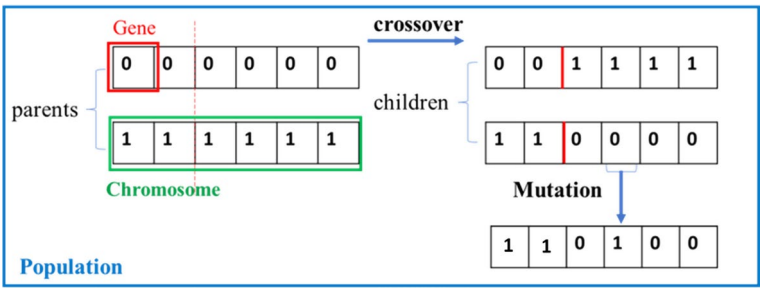

Fig. 4 Display of the chromosome, gene, population and crossover and mutation operators in GA algorithm

2.2.4 Ant colony optimization (ACO)

We use the method described by Chaudhuri et al. (2014). A meta-heuristic algorithms usually take inspiration from social behaviors of animals. The ACO for solving combinatorial optimization problems was first used to solve the traveling salesman problem (TSP): given that one single traveling salesman must visit a set of n cities, how to minimize the total distance length, which is a well-known NP-hard problem.

In ACO algorithm, the ants work together to find the shortest path between the nest and the food source so that they can transport food to the nest in as little time as possible. During their trip's, ants leave a chemical trail called pheromone on the ground. The high number of traffic and the creation of more pheromones leads to an optimal path. The pheromone evaporation and the probability-accident allow the ants to find the shortest path. The ACO algorithm manages the scheduling of three activities. The first step consists mainly of initialization of the pheromone trail. In the iteration (second) step, each ant constructs a complete solution to the problem, according to a probabilistic state transition rule. The state transition rule depends mainly on the state of the pheromone. The third step updates the quantity of pheromone. A global pheromone updating rule is applied in two phases: in the first phase, a fraction of the pheromone evaporates; and in the second phase, each ant deposits an amount of pheromone which is proportional to the fitness of its solution. This process is iterated until a stopping criterion is met (Chaudhuri et al., 2014).

2.3 Used software

Data analysis was performed using MATLAB and R Studio software. I used the MAS, car, boot, spatial, lmtest, and survival packages for regression modeling and the e1071 package for SVR modeling. The "nftool" package in MATLAB software (2017) was employed for the MLP model of the ANN algorithm. Also, GA and ACO optimization were performed in MATLAB software (2017).

3 Results

The present study aimed to estimate and impute the missing data of the five stations in Iran regarding the long-term historical monthly rainfalls with the missing values. The imputation of the missing values before the statistical analysis of the meteorological data could definitely decrease uncertainty and provide robust results. To this end, the classic parametric multiple regression method was compared with nonparametric, meta-heuristic methods. Missing percentage of this stations is listed in the last column of Table 1. The missingness mechanism for each of these five stations was MAR (Missing at Random) (Farzandi, 2019).

Some stations in the neighboring countries of Iran have available long-term data and are located in Turkmenistan, Azerbaijan, Uzbekistan, Pakistan, Oman, the United Arab Emirates, Qatar, Iraq, and Armenia (Fig. 1). The available stations in the neighboring countries ($n=81$) were surveyed. According to the results of the statistical analysis, most of the stations were not reliable in terms of missing data management. Therefore,

reliable stations were selected based on the two factors of closeness and complete dataset (Fig. 1).

Initially, the regression procedure was adopted, followed by the use of GA and ACO, to improve the adequacy of the estimation of the missing values using regression. Moreover, ANN and SVR were used separately to determine the missing values.

3.1 Regression

The regression method on this paper prepares an initial model for GA and ACO. Because the evolutionary methods (GA and ACO) require an initial pattern. After examining different regression patterns, the best model for each station was selected according to Eqs. 2-6 for fitting GA and ACO. The monthly precipitation in Ashgabat (R_{Ash} as X_1), Sarakhs (R_{Ser} as X_2), and Kooshka (R_{Kus} as X_3) was selected as the predictive variables for the management of the missing data of Mashhad (R_{Mas} as Y) (Eq. 2). The monthly precipitation in Baghdad (R_{Bag} as X_1) and Lenkoran (R_{Len} as X_2) was selected for Tehran (R_{Teh} as Y) (Eq. 3). The monthly precipitation of Basra (R_{Bas} as X_1) and Diwaniya (R_{Diw} as X_2) was selected for Isfahan (R_{Esf} as Y) (Eq. 4). The monthly precipitation of Bahrain (R_{Bah} as X_1) and Sharjah (R_{Sha} as X_2) was selected for Jask (R_{Jas} as Y) (Eq. 5), and the monthly precipitation of Bahrain (R_{Bah} as X_1) was chosen for Bushehr (R_{Bus} as Y) (Eq. 6).

$$R_{Mas} = \beta_0 + \beta_1 R_{Ash} + \beta_2 R_{Kus} + \beta_3 R_{Ser} \tag{2}$$

$$R_{Teh} = \beta_0 + \beta_1 R_{Bag} + \beta_2 R_{Len} \tag{3}$$

$$R_{Esf} = \beta_0 + \beta_1 R_{Bas} + \beta_2 R_{Diw} \tag{4}$$

$$R_{Jas} = \beta_0 + \beta_1 R_{Bah} + \beta_2 R_{Sha} \tag{5}$$

$$R_{Bus} = \beta_0 + \beta_1 R_{Bah} \tag{6}$$

Initially, the missing data of each station were restored by the fitting of several multiple linear regression models to the monthly precipitations. It is notable that in this analysis, the outlier data were eliminated, and the best-fitted model was selected for the five stations separately (Table 2). The adjusted coefficient of variation (R^2) was within the range of 0.42–0.85, which was considered adequate (Fox, 2016). The optimal results of the selected models are presented in Table 2. The P values of β_0–β_3 indicated that these parameters were significant at 99%. In addition, all the variance inflation factors were less than 3.1 (Table 2).

The RMSE of the five patterns was within the range of 7.8–20.9 mm. Moreover, the statistics of the F-test were adequately significant in all the models (range: 222–845). The Durbin–Watson test statistics (D–W) were within the unmatched limits of the Durbin–Watson table (1.5–2.5), which confirmed the independence of the residual (Table 2).

According to the Chi-square test results in the non-constant of variance (NCV) test, the variance of residuals was not stable in all the stations. The NCV value of the stations in Mashhad, Isfahan, Tehran, Jask, and Bushehr was estimated at 273, 108, 114, 408, and 407, respectively, with all the P values close to zero. In addition, the range of the average cook-distance statistics in all the patterns was 0.0024–0.025, which indicated the lack of

an outlier. Shapiro–Wilk test showed that the normality assumption of the residuals was rejected. But due to a large amount of data, we ignored it.

3.2 Optimization of the patterns with GA and ACO

Due to the lack of basic assumptions, the results of the regression are not invoked and non-parametric methods (GA and ACO) are chosen to optimize the patterns.

GA and ACO make slight changes to its solutions slowly until getting the best solution. Here, the objective function (optimizer) is to minimize the amount of error (RMSE) and the objective function (optimizer) is to minimize the amount of error (RMSE) as well. That is, it changes the coefficients of regression patterns so that the error is minimized.

The regression parameters were estimated using the least-squares error method. GA and ACO were the optimization methods that could improve the accuracy of parameters β_0, β_1, β_2, and β_3 in Eqs. 2–6. After estimating the new coefficients, the RMSEs of the latest and previous patterns (i.e., regression patterns) were compared.

GA requires some default parameters, and the coefficient range was selected to be -20 to $+20$ in the present study based on the pilot analysis. Parameters estimate in Table 2 shows that the regression coefficients were at least 0.1 and at most 7.5. So coefficients less than -20 and more than $+20$ are inconceivable.

The maximum iteration was within the range of 200–1000 in each pattern. The assumptions and initial inputs for the implementation of GA included the initial population size of 30, the mutation parameter of 0.02, gamma of 0.05, number of parents of two, the mutation rate of 0.3, and crossover rate of 0.8 with the roulette wheel as the selected method. The output results are shown in Table 3.

ACO requires initial values and constant parameters. The initial population count was 10, with the sample size of 50, the deviation-distance ratio of one, and intensification factor (q) of 0.5. The final results of GA and ACO based on the optimized parameters and

Table 2 Results of fitting regression patterns for the 5 stations rainfall, include: Coefficients, Variance Inflation factor (VIF), Durbin–Watson Statistics, RMSE (mm), and Pattern power (F-statistic)

Stations name	Parameter estimate				VIF	R_{adj}^2	$RMSE_{reg}$ (mm)	F-statistic	D–W
	β_0	β_1	β_2	β_3					
Mashhad	**2.05**	**0.29**	**0.11**	**0.68**	< 3.1	0.78	9.79	968	1.76
p value	$1.1*10^{-4}$	$2.0*10^{-16}$	$4.1*6^{-9}$	$2.0*10^{-16}$					
Isfahan	**2.92**	**0.36**	**0.19**		< 1.6	0.57	7.89	222	1.91
p value	$3.5*10^{-8}$	$2.0*10^{-16}$	$1.2*10^{-8}$	–					
Tehran	**6.32**	**0.71**	**0.02**		< 1.1	0.46	13.43	294	1.91
p value	$3.3*10^{-15}$	$2.0*10^{-16}$	$3.0*10^{-4}$	–					
Jask	**0.18**	**0.36**	**0.70**		< 1.8	0.85	6.56	845	1.88
p value	$6.0*10^{-2}$	$1.9*10^{-13}$	$2.0*10^{-16}$	–					
Bushehr	**7.48**	**1.48**			–	0.42	20.96	734	1.50
p value	$2.0*10^{-16}$	$2.0*10^{-16}$	–	–					

Pattern of Mashhad in first row has three independent variables (β_0–β_3 are the estimated coefficients). Isfahan and Tehran in the second and third row has two independent variables (β_0–β_2 are the estimated coefficients). Bushehr in the fourth row has one independent variables (β_0–β_1 are the estimated coefficients). (–) show that there is no second or third independent variable. β_0 is the width of the origin

Methods are compared with the error value (RMSE)

Table 3 Optimized parameters of patterns using the GA & ACO and RMSE

Stations name	Parameter estimate by GA				$RMSE_{GA}$ (mm)	Parameter estimate by ACO				$RMSE_{ACO}$ (mm)
	β_0	β_1	β_2	β_3		β_0	β_1	β_2	β_3	
Mashhad	-0.0026	0.28	0.12	0.67	**2.560**	-2.77×10^{-7}	0.26	0.14	0.65	**2.559**
Isfahan	8.17×10^{-3}	0.39	0.27	–	**2.512**	1.11×10^{-13}	0.37	0.30	–	**2.513**
Tehran	0.95	0.80	0.02	–	**3.498**	0.95	0.80	0.02	–	**3.498**
Jask	-2.99×10^{-10}	0.21	0.81	–	**2.485**	-3.57×10^{-12}	0.2	0.82	–	**2.484**
Bushehr	1.41×10^{-3}	1.86	–	–	**4.021**	9.09×10^{-13}	1.86	–	–	**4.021**

Pattern of Mashhad in first row has three independent variables (β_0–β_3 are the estimated coefficients). Isfahan and Tehran in the second and third row has two independent variables (β_0–β_2 are the estimated coefficients). Bushehr in the fourth row has one independent variables (β_0–β_1 are the estimated coefficients). (–) show that there is no second or third independent variable. β_0 is the width of the origin

Methods are compared with the error value (RMSE)

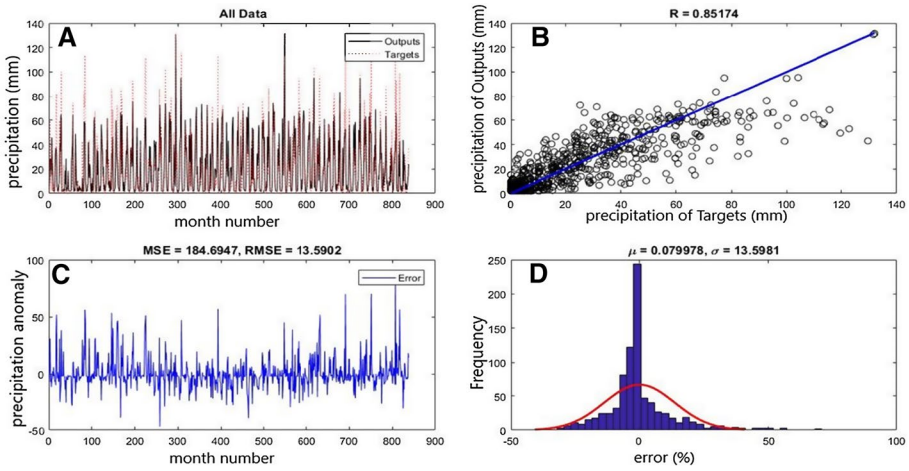

Fig. 5 Time series of the observed monthly rainfall (1893–2017) of Mashhad (Targets) and predicted data by a network (Outputs) in mm (**A**), determination coefficient between Targets and Output (**B**), time series of error values (**C**), and μ distribution (**D**)

RMSEs of the studied stations are shown in Table 3. According to the information in this table, the RMSE of all the patterns was less than 4 mm. Furthermore, no significant difference was observed between GA and ACO in terms of the RMSE.

According to the information in Table 2, RMSE was greater than 6.65 mm, while the minimum RMSE of ACO and GA was 2.48 mm, indicating a difference between the regression models, GA, and ACO in this regard. Therefore, it could be concluded that evolutionary methods could significantly reduce the error of the regression patterns.

3.3 Modeling with ANN

ANN is a flexible mathematical structure, which is capable of identifying the complex, nonlinear correlations between the input and output datasets. In the current research, the MLP model in ANN was selected and implemented to the model precipitation data of the stations to predict the missing values. According to the findings, 70% of the data were on training, 15% were on validation, and 15% were on randomized testing.

The number of the hidden layers was set at three based on trial and error. The *tanh*[1] sigmoid function was considered optimal, and the Loewenberg Marquart represented the training algorithm. Data were split randomly, and the performance level was measured based on the MSE. The model was allowed to repeat 500 times. The efficiency (error rate) and number gradient were close to zero, with 20 allowed failures. Notably, the occurrence of any of these factors disrupts the process. The result of all output of ANN was huge in all stations, so do not report them.

Figures 5, 6, 7 show the studied samples. As depicted, the graphs of the observed monthly precipitation in Mashhad (targets) and predicted data (outputs) indicated that the

[1] Hyperbolic tangent.

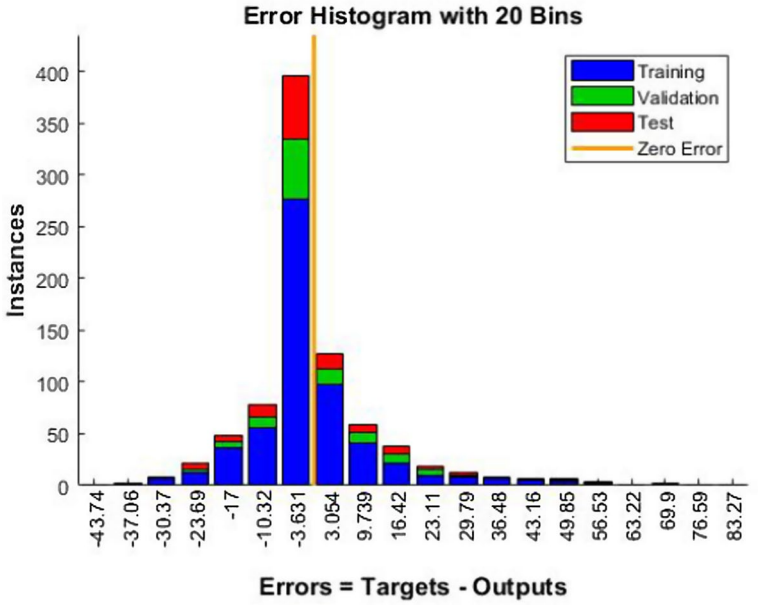

Fig. 6 Distribution of errors (Mashhad monthly precipitations). Training (blue), validation (green), and experiment (red)

model output could partly justify the monthly rainfall patterns, and the determination coefficient of data was estimated at 0.852 (Fig. 5).

Time series of the observed monthly rainfall (From January (1893) to December (2017) of Mashhad (Targets) and predicted data by a network (Outputs) in mm is shown at the top left of Fig. 5. As can be seen, the time series of the error values indicated the RMSE was 13.59 mm (Fig. 5). Besides, the error histogram (μ distribution) is illustrated at the down-right of Fig. 5. This indicated the distribution of the errors with a little skewness, which is due to some extreme amounts of precipitation.

The stacked distribution of the modeling errors is shown in Fig. 6, and education (blue), validation (green), and experiment (red) are depicted in Fig. 6. Figure 7 shows the linear regression and determination coefficient between the predicted and observed values of the monthly precipitation of Mashhad with the neural network in terms of training, validation, test, and all the data. Accordingly, the determination coefficient for training, validation, test, and all the data was 0.858, 0.849, 0.831, and 0.852, respectively (Fig. 7).

3.4 Support vector regression

The SVR model was fitted for the prediction of the missing data of monthly rainfall in the five stations. The radial kernel function was selected for the SVR algorithm. The input parameters of cost and epsilon in the first stage were considered to be 1 and 0.1, respectively. For instance, in the final stage (four repetitions), the best-estimated parameter for rainfall in Mashhad was achieved by the cost of 1.4 and epsilon of 0.1 in the fourth stage. In addition, the number of support vectors was 382, with the RMSE of 11.88 mm (Table 4). The RMSE value obtained from the ANN and SVR is presented in Table 5.

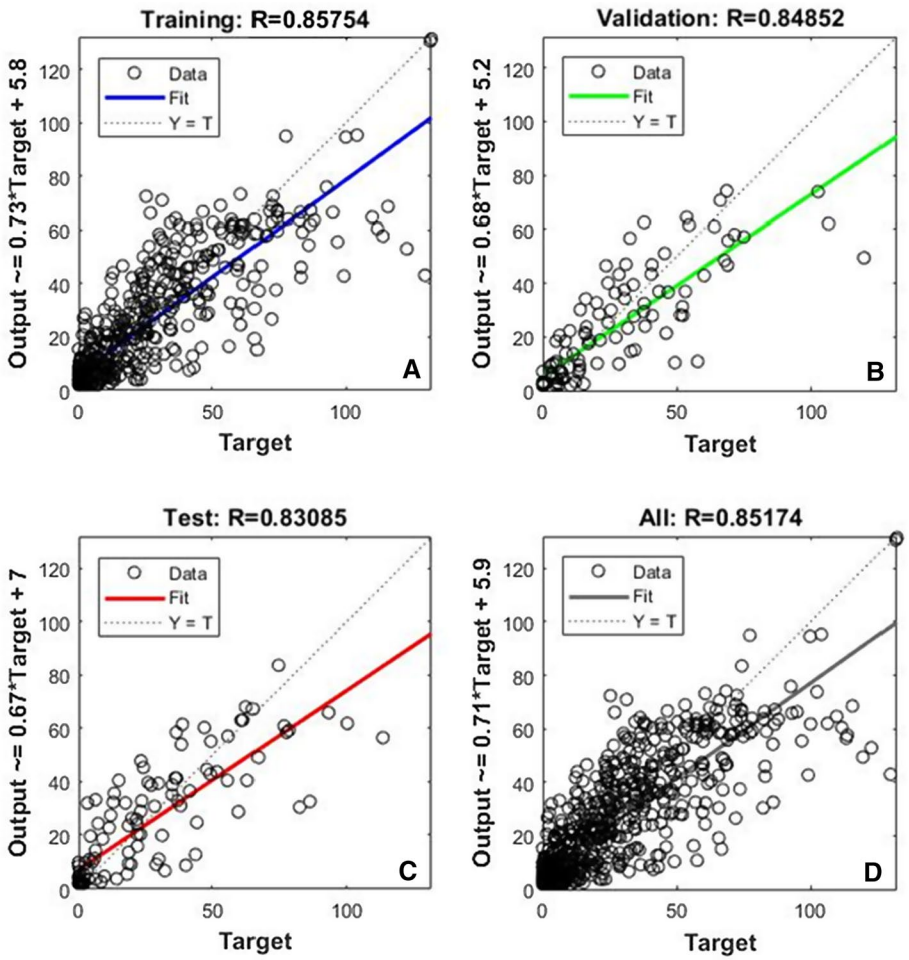

Fig. 7 Draw the predicted values against the observed values (Mashhad monthly precipitations (mm)) by the neural network in the training (**A**), validation (**B**), test (**C**), and all data (**D**)

In the final comparison, the RMSEs of the regression, GA, ACO, ANN, and SVR methods were calculated (Table 5). Accordingly, the RMSEs of GA and ACO were the same and minimum, while the RMSEs of SVR and ANN were the maximum. Therefore, it could be concluded that the monthly missing data of the selected stations were imputed with an optimized pattern of ACO or GA. In the present study, ACO required less repetition to achieve convergence, while it might have underestimated some of the missing values, especially in the stations in Tehran and Mashhad. As a result, the missing values were imputed using the GA for the stations in Tehran and Mashhad, while the data of the other stations were imputed using the ACO algorithm. It is also notable that the mentioned analyses were performed on a monthly scale, and the rainfall series was restored on a monthly scale as well. The total annual precipitation per year calculated by the sum of monthly precipitation values recorded in that year.

Table 4 Modified coefficients of precipitation patterns with SVR algorithm

step	Best cost	Best epsilon	Gama	Weight (W)			b	Number of Support Vectors	$RMSE_{SVR}$ (mm)
				R_{Ash}	R_{Ser}	R_{Kus}			
1	1	0.1	0.167	13.97	19.04	13.30	−1.24	381	12.18
2	1	0.1	0.167	13.97	19.04	13.30	−1.24	381	12.18
3	1.4	0.16	0.167	16.23	20.84	14.30	−1.32	328	11.89
4	1.4	0.1	0.167	16.36	20.28	14.12	−1.31	382	11.88

Table 5 RMSE (mm) value obtained from fitted 5 methods (regression, GA, ACO and modeling by ANN and SVR)

Stations name	RMSE (mm)				
	Regression	GA	ACO	ANN	SVR
Mashhad	9.8	2.6	2.6	13.6	11.8
Isfahan	7.9	2.5	2.5	6.7	7.6
Tehran	13.4	3.5	3.5	15.4	15.8
Jask	6.6	2.5	2.5	16.5	11.9
Bushehr	21.0	4.0	4.0	34.2	35.0

Figure 8 shows the completed annual precipitation time series of the five selected stations. Figure 8A was related to 125 years of annual precipitation in Mashhad, 7B was related to 125 years of precipitation in Isfahan, 7C was referred to 137 years of precipitation in Tehran, D was linked to 125 years of precipitation in Jask, and E related to 140 years of precipitation in Bushehr. In Fig. 8 the black series were related to long-term time-series precipitation at the five stations in Iran. The missing data imputed by GA or ACO methods are shown by red color. Also, the average annual precipitation line was plotted in each series. The main stages of this research are shown in the flow chart (Fig. 9).

4 Discussion

Environmental models typically require a complete time series of meteorological inputs; thus, reconstructing missing data are a crucial issue in the functionality of such physical and statistical models. Estimating with missing values is bias. So, determining missing data must be as precise as possible.

In the present study, we implemented and compared several methods to improve the accuracy of estimating missing values. These methods include classic approaches (multiple regression), artificial intelligence (ANN and SVR), and evolutionary methods (GA and ACO) to repair missing precipitation data of the five stations in Iran. First, several regression models fitted to data for estimating missing values. If the data have some hypotheses such as normality, independent of errors, variance stability, then the estimation of parameters is Best Linear Unbiased Estimator (BLUE) with minimum variance (Fox, 2016). Monthly rainfall data are not normal and have deviated from this hypothesis (noisy data). For improving the accuracy of regression parameters, we used GA and ACO optimizations methods. According to the results (Table 5), GA and ACO could enhance the accuracy of estimating the regression parameters of our monthly precipitation data. Therefore, this

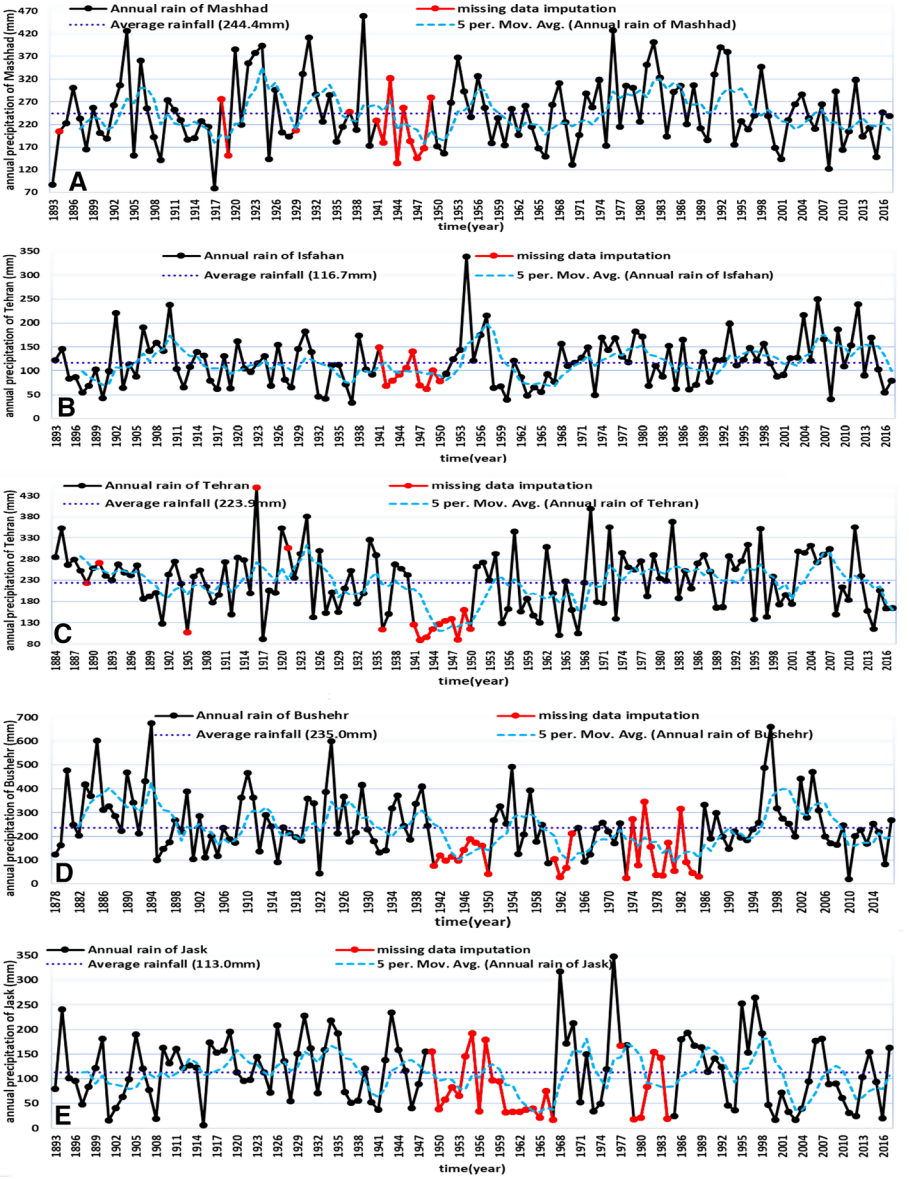

Fig. 8 The completed time series of annual precipitation for the five stations. The black series were related to long-term precipitation. Missing data imputed by GA or ACO methods are shown by red color. The average annual precipitation line was plotted in each series. **A** Mashhad, **B** Isfahan, **C** Tehran, **D** Bushehr, and **E** Jusk stations

study found that GA and ACO optimizations have greatly enhanced the accuracy of missing data estimation. This result is consistent with the application of evolutionary methods. Because the main motivation behind using evolutionary algorithms in data mining is their attractive features that enable them to resolve some of the drawbacks in conventional

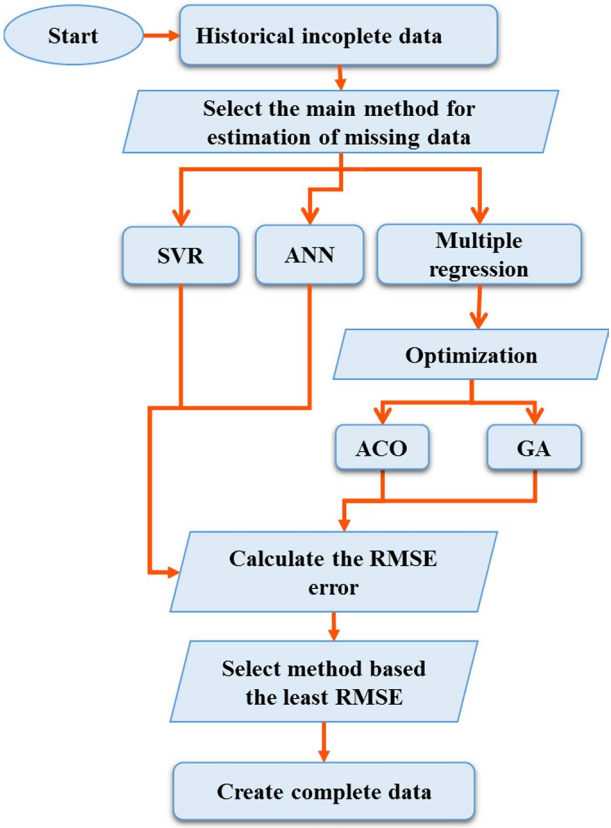

Fig. 9 Flowchart of the main stages for imputation of missing data in this research

data mining techniques and allow them to discover novel solutions, such as their robustness when dealing with noisy data, and their ability to interpret data without any a priori knowledge (Abbass et al., 2002).

The authors of this article in another study found that the machine learning methods (SVR & ANN) significantly enhanced the accuracy of the missing temperature data estimation and performed better than traditional and evolutionary approaches (Farzandi et al., 2019). Therefore, evolutionary methods are the best for estimate noisy data such as precipitation, and machine learning methods (ANN& SVR) are more suitable for high correlation data (with independent variables), such as temperature.

5 Conclusion

The current research aimed to predict the missing values of the five stations in Iran (Table 1, Fig. 1) regarding the long-term monthly precipitation (125–140 years). Several methods were used to this end, including regression, GA, ACO, ANN, and SVR.

Comparison of RMSE of the mentioned methods indicated that evolutionary methods (GA and ACO) could better estimate the missing monthly precipitation data. The RMSEs

of GA and ACO in the selected stations were within the range of 2.6–4.0 mm. On the other hand, machine learning methods (ANN and SVR) could not increase the accuracy of imputing the missing data of monthly precipitation compared to the regression method (Fig. 9). As a result, the complete historical annual precipitation of the five stations in Iran became available (Fig. 8). Our findings could be valuable in addressing issues such as water resources, the return of extreme precipitation, droughts, climate changes, and global warming.

References

Abbass, H. A., Sarker, R. A., & Newton, C. S. (2002). *Data mining, a heuristic approach-IGI global* (p. 300). Idea Group Publishing.

Aguilera, H., Carolina, G. A., & Carmen, S. H. (2020). Estimating extremely large amounts of missing precipitation data. *Journal of Hydroinformatics,* 22(3): 578–592.

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression. *Information Sciences*, *233*(1), 25–35.

Belala, F., Hirche, A. D., Muller, S., Tourki, M., Salamani, M., Grandi, M., Ait Hamouda, T., & Boughani, M. (2018). Rainfall patterns of Algerian steppes and the impacts on natural vegetation in the 20th century. *Journal of Arid Land, 10*(4), 561–573.

Chaudhuri, S., Goswami, S., Das, D., & Middey, A. (2014). Meta-heuristic ant colony optimization technique to forecast the amount of summer monsoon rainfall: Skill comparison with Markov chain model. *Theoretical and Applied Climatology, 116*(3–4), 585–595.

Coulibaly, P., & Evora, N. D. (2007). Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology, 341*(1–2), 27–41.

Dastorani, M. T., Moghadamnia, A., Piri, J., & Rico-Ramirez, M. (2010). Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment., 166*(1–4), 421–434.

Escalante-Sandoval, C., & Nuñez-Garcia, P. (2017). Meteorological drought features in northern and northwestern parts of Mexico under different climate change scenarios. *Journal of Arid Land, 9*(1), 65–75.

Farzandi, M. (2019). The Hybrid EM and evolutionary algorithms for estimating and analyzing missing data in meteorology Case study: 130-years monthly precipitation and temperature of the five stations in Iran. PhD dissertation, Ferdowsi University of Mashhad, Iran.

Farzandi, M., Sanaeinejad, H., Ghahraman, B., & Sarmad, M. (2019). Imputation of missing meteorological data with evolutionary and machine learning methods, case study: long-term monthly precipitation and temperature of Mashhad. *Journal of Water and Soil, 33*(2), 361–377.

Fox, J. (2016). *Applied regression analysis and generalized linear models* (p. 816). SAGE Publications, Inc.

Geiß, C., Jilge, M., Lakes, T., & Taubenböck, H. (2016). Estimation of seismic vulnerability levels of urban structures with multisensor remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9*(5), 1913–1936.

Golkar Hamzee Yazd, H. R., Salehnia, N., Kolsoumi, S., & Gerrit, H. (2019). Prediction of climate variables by comparing the k-nearest neighbor method and MIROC5 outputs in an arid environment. *Climate Research, 77*, 99–114.

http://sdwebx.worldbank.org. (https://climateknowledgeportal.worldbank.org/)

https://climexp.knmi.nl. (Koninklijk Nederlands Meteorologisch Institute Climate Explorer)

Islamic Republic of Iran Meteorological Organization (https://www.irimo.ir/eng/index.php)

Iqbal, M., Wen, J., Wang, Sh., & Adnan, M. (2018). Variations of precipitation characteristics during the period 1960–2014 in the Source Region of the Yellow River, China. *Journal of Arid Land, 10*(3), 388–401.

Jacob, D., Reed, D. W., & Robson, A. J. (1999). *Choosing a pooling group. Flood estimation handbook* (Vol. 3). Institute of Hydrology.

Kazemzadeh, M., & Malekian, A. (2018). Homogeneity analysis of streamflow records in arid and semi-arid regions of northwestern Iran. *Journal of Arid Land, 10*(4), 493–506.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (p. 408). Wiley.

Miang Kueh, S., & Kuok Kuok, K. (2016). Precipitation downscaling using the artificial neural network BatNN and development of future rainfall intensity-duration-frequency curves. *Climate Research, 68*, 73–89.

Patil, D. V., Bichkar, R. S. (2010). Multiple imputation of missing data with genetic algorithm based techniques. IJCA special issue on evolutionary computation for optimization techniques.

Ramesh, S. V. T., Tufail, M., & Ormsbee, L. (2009). Optimal functional forms for estimation of missing precipitation data. *Journal of Hydrology*, *374*(1–2), 106–115.

Salehnia, N., Alizadeh, A., Sanaeinejad, H., Bannayan, M., Zarrin, A., & Hoogenboom (2017). Estimation of meteorological drought indices based on AgMERRA precipitation data and station-observed precipitation data. *Journal of Arid Land*, *9*(6), 797–809.

Salehnia, N., Salehnia, N., Ansari, H., Kolsoumi, S., & Bannayan, M. (2019). Climate data clustering effects on arid and semi-arid rainfed wheat yield: A comparison of artificial intelligence and K-Means approaches. *International J. of Biometeorology, 63*(7), 861–872. https://doi.org/10.1007/s00484-019-01699-w

Sattari, M., Rezazadeh-Joudi, A., & Kusiak, A. (2017). Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research*, 48(4), 1032–1044.

Serrano-Notivoli, R., de Luis, M., Ángel Saz, M., & Beguería, S. (2017). Spatially based reconstruction of daily precipitation instrumental data series. *Climate Research, 73*, 167–186.

Seyyednezhad Golkhatmi, N., Sanaeinejad2, S. H., Ghahraman, B., Rezaee Pazhand, H. (2012). Extended modified inverse distance method for interpolation rainfall. International Journal of Engineering Inventions, 1(3): 57-65.

Smithsonian Institution. (1927, 1934, 1947): World weather records, 1910–1920, 1921–1930, 1931–1940. Smithson. Miss C. Collect. 79,90,105. (Publication 2913, 3216, 3803)

Smola, A. J., & Vishwanathan, S. V. N. (2008). *Introduction to machine learning* (p. 234). Cambridge University Press.

Türkeş, M., Yozgatlıgil, C., Batmaz, I., İyigün, C., Kartal Koç, E., Fahmi, F. M., & Aslan, S. (2016). Has the climate been changing in Turkey? Regional climate change signals based on a comparative statistical analysis of two consecutive time periods, 1950–1980 and 1981–2010. *Climate Research, 70*, 77–93.

Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I. (2013). Comparison of missing value imputation methods in time series: The case of Turkish meteorological data. *Theory Apply Climatology, 112*(1–2), 143–167.