CrossMark

New Approaches for Estimation of Monthly Rainfall Based on GEP-ARCH and ANN-ARCH Hybrid Models

Saeid Mehdizadeh[1] · Javad Behmanesh[1] ·
Keivan Khalili[1]

Abstract Accurate estimation of rainfall has an important role in the optimal water resources management, as well as hydrological and climatological studies. In the present study, two novel types of hybrid models, namely gene expression programming-autoregressive conditional heteroscedasticity (GEP-ARCH) and artificial neural networks-autoregressive conditional heteroscedasticity (ANN-ARCH) are introduced to estimate monthly rainfall time series. To fulfill this purpose, five stations with various climatic conditions were selected in Iran. The lagged monthly rainfall data was utilized to develop the different GEP and ANN scenarios. The performance of proposed hybrid models was compared to the GEP and ANN models using root mean square error (RMSE) and coefficient of determination (R^2). The results show that the proposed GEP-ARCH and ANN-ARCH models give a much better performance than the GEP and ANN in all of the studied stations with various climates. Furthermore, the ANN-ARCH model generally presents better performance in comparison with the GEP-ARCH model.

Keywords Estimation · Rainfall · GEP-ARCH · ANN-ARCH

1 Introduction

Rainfall is one of the most important parameters in the hydrologic cycle. Since Iran is located in an arid and semi-arid climatic region, the estimation of rainfall plays an important role to

✉ Javad Behmanesh
   j.behmanesh@urmia.ac.ir

   Saeid Mehdizadeh
   saied.mehdizadeh@gmail.com

   Keivan Khalili
   khalili2006@gmail.com

[1]   Department of Water Engineering, Urmia University, Urmia, Iran

🦋 Springer

plan and manage water resources. Moreover, accurate estimation of rainfall is vital in the assessment of hydrological parameters such as runoff, flood, sediment, drought, agriculture, irrigation scheduling, groundwater, etc. Nevertheless, accurate estimation of rainfall is difficult due to its spatial and temporal variations (Kisi and Cimen 2012). Although the estimation of rainfall is not impossible, it faces great complexities due to its non-linear behavior.

Artificial neural networks (ANN) and gene expression programming (GEP) are considered as soft computing approaches which are extensively used to model hydrologic variables. The use of ANN is more common than GEP for estimating the rainfall. However, the GEP has widely been used in other engineering fields (Marti et al. 2013; Imani et al. 2014; Ebtehaj et al. 2015; Gocic et al. 2015; Yassin et al. 2016; Mehdizadeh et al. 2016, 2017; Shiri et al. 2017; Behmanesh and Mehdizadeh 2017). Ramirez et al. (2005) applied the ANN to forecast the rainfall in Sao Paulo, Brazil. The ANN, as a result, was superior to linear regression model. Partal and Kisi (2007) employed a conjunction model, namely wavelet-neuro fuzzy for prediction of daily precipitation in three stations in Eagean region, Turkey. The results showed that the proposed conjunction model performs better than the classical neuro-fuzzy model. Wu et al. (2010) predicted monthly and daily rainfall time series using modular ANN coupled with data-preprocessing methods in China and India. The results revealed that the benefits of modular ANN than other models are significant, especially in the prediction of daily rainfall. Moustris et al. (2011) predicted the precipitation using ANN at four meteorological stations in Greece. Their findings confirmed the satisfactory of precipitation forecasts by the ANN. Kashid and Maity (2012) estimated summer monsoon monthly rainfall on homogenous regions of India using genetic programming (GP). Mekanik et al. (2013) employed the ANN and multiple regression (MR) to forecast seasonal spring rainfall in Victoria, Australia. The outcomes revealed the superiority of the ANN over MR. Venkata Ramana et al. (2013) used wavelet neural networks to estimate monthly rainfall in Darjeeling rain gauge station. The wavelet neural networks models were found to perform better than the ANN models. Abbot and Marohasy (2014) used ANN to predict monthly rainfall as a continuous variable in Queensland, Australia. The results indicated the usefulness of climate indices for the estimation of rainfall. He et al. (2015) introduced hybrid wavelet neural network model with mutual information and particle swarm optimization to forecast the monthly rainfall in Australia. The results demonstrated that the proposed model increased the accuracy of the prediction of monthly rainfall compared to the reference models. Shenify et al. (2016) compared the performance of wavelet transform-support vector machine (WT-SVM) with GP and ANN for the estimation of monthly precipitation in Serbia. Consequently, the WT-SVM estimations were superior to the GP and ANN models.

Beside soft computing methods, there are several time series models to estimate the time series of hydrological variables such as rainfall. Chinchorkar et al. (2012) stated that the rainfall is often predicted using stochastic models due to the random properties of rainfall time series. Some of these models are autoregressive (AR), autoregressive moving average (ARMA), etc. Autoregressive conditional heteroscedasticity (ARCH) model which was presented by Engle (1982) performs as a non-linear model. The variance in this model is not considered as a constant overtime.

In fact, time series of hydrologic variables such as rainfall have composed of two components; deterministic and random parts. Literature review reveals that the previous conducted studies about the rainfall forecasting have focused only on the deterministic part of rainfall time series. This means that the random component of rainfall data has not been considered in previous studies. Herein, random component was considered in the modeling process. The

present study attempts to create novel methods by combining the ARCH time series model with the GEP and ANN to enhance the estimation accuracy of the rainfall time series at different climatic regions in Iran. In the present study, the deterministic part of rainfall time series was estimated using the GEP and ANN models whereas the random part was determined by the ARCH model.

2 Materials and Methods

2.1 Study Area and Data Used

Iran, with an area of over 1,648,000 km^2, is located in the southwest of Asia. About 94.8% of Iran has arid or semi-arid climate with low amounts of rainfall and high evapotranspiration (Khalili et al. 2016). In the present study, five regions with various climatic conditions were chosen. It should be noted that the aridity index defined by United Nations Environment Program (UNEP 1992) was employed to classify the selected stations in terms of climatic aspect (Ashraf et al. 2014). Table 1 presents the geographical and climatic information of the studied stations. Also, Fig. 1 demonstrates the spatial distribution of the considered stations in Iran.

The required data, i.e. the monthly rainfall of the studied stations was obtained from the Islamic Republic of Iran Meteorological Organization (IRIMO) during 1981 to 2012. Here, the whole data was divided into two parts; training and testing datasets. Around 75% (1981–2004, i.e. 288 months) and 25% (2005–2012, i.e. 96 months) of the data were utilized in training and testing stages, respectively. Figure 2 depicts the time series of the monthly rainfall for the studied stations during 1981 to 2012. Descriptive statistics of the monthly rainfall data for the study regions are presented in Table 2. As seen, humid and hyper-arid stations; i.e. Rasht and Zahedan have the maximum and minimum values of standard deviation, respectively. On the other hand, the lowest and highest skewness are related to Gorgan and Zahedan, respectively.

2.2 Artificial Neural Networks

Artificial neural networks (ANN) were created on the basis of the extensive internal communications like the human brain and nervous system. The ANN is presented as a structure like the biological one in the human brain. In these models, at first, the data is introduced to the model. Then, the training process will be done and the obtained (trained) model will be used for the next steps. The general structure of the ANN is composed of three layers;

Table 1 Geographical and climatic information of the studied stations

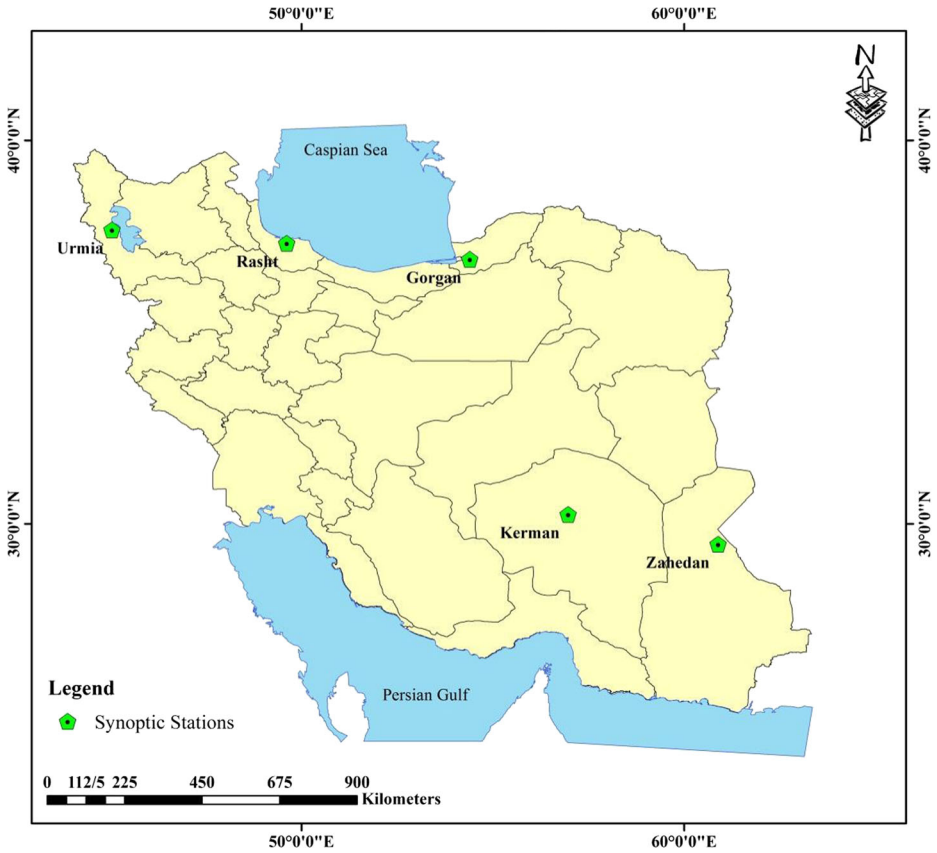| Station	Latitude (N)	Longitude (E)	Elevation (m)	Climate type
Rasht	37° 19'	49° 37'	−8.6	Humid
Gorgan	36° 54'	54° 24'	0.0	Sub-humid
Urmia	37° 40'	45° 03'	1328.0	Semi-arid
Kerman	30° 15'	56° 58'	1753.8	Arid
Zahedan	29° 28'	60° 53'	1370.0	Hyper-arid

Fig. 1 The location of studied stations in Iran

1- Input layer: In this layer, the input data is introduced to the model.
2- Hidden layer: In the hidden layer, the input data is processed.
3- Output layer: The results of the model are produced in the output layer.

Each layer consists of some neurons. Each neuron receives information from a series of input data, and after processing, it delivers the results to the output. Finally, the output of the mentioned neuron is used as the next layer input. The number of neurons in the input and output layers are determined in accordance with the nature of the studied phenomenon. In this study, the lagged monthly rainfall data were considered as input variables, while the output of the ANN was defined as the rainfall of current month. Moreover, the number of neurons in the hidden layer is specified by user using trial and error procedure in order to reduce the error rate of the network. In the present study, the optimum number of neurons in the hidden layer changes between 1 and 6 for the different applied scenarios and considered stations. The neurons of each layer are connected to the neurons of next layer by weights. Beside weights, the biases are the elements of the hidden and output layers. During the training process, the weights and biases are continuously changed until the best structure of the network is found.
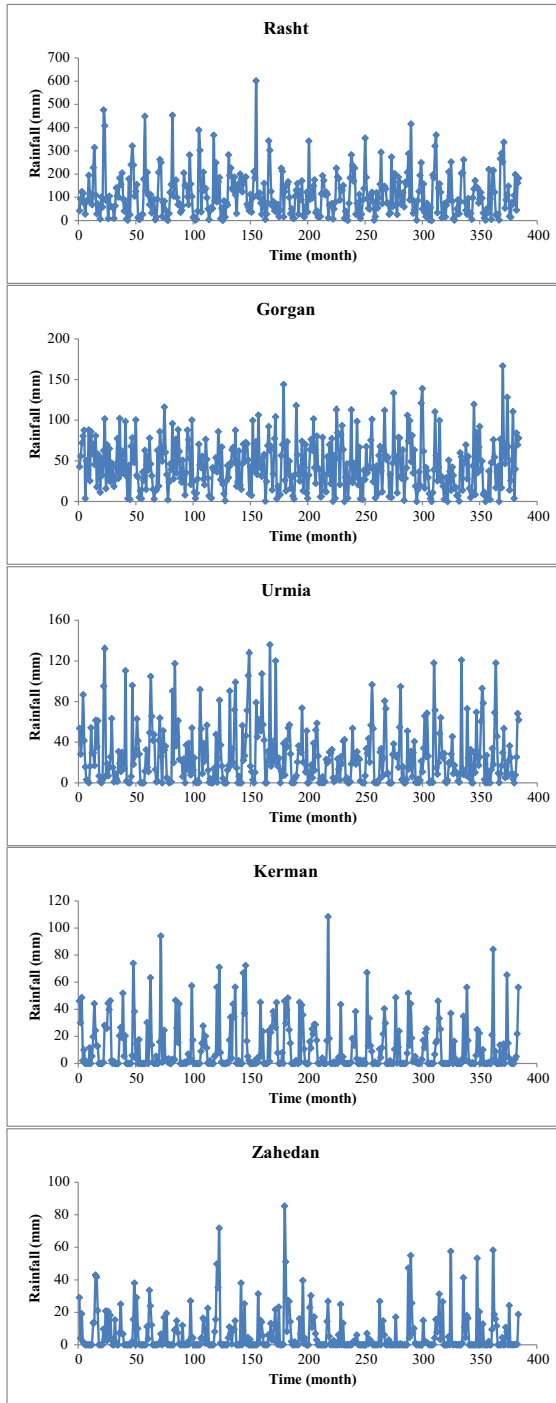
Fig. 2 Time series of the monthly rainfall for the studied stations during 1981–2012

Table 2 Statistical properties of the monthly rainfall time series

Station	Minimum (mm)	Maximum (mm)	Mean (mm)	Standard deviation (mm)	Skewness
Rasht	0.0	601.4	112.0	93.3	1.46
Gorgan	0.0	166.8	44.7	30.9	0.74
Urmia	0.0	136.0	26.3	28.4	1.52
Kerman	0.0	108.4	11.3	18.0	2.07
Zahedan	0.0	85.4	6.5	12.1	2.83

Furthermore, transfer functions are used to create a relationship between input and output parameters. In this study, tangent-sigmoid and linear transfer functions were used in the hidden and output layers, respectively. More details about the ANN can be found in Haykin (1998).

The ANN program was written in MATLAB programming language. In the present study, three-layer feed forward neural networks were used to estimate the monthly rainfall time series in the studied area. Moreover, Levenberg-Marquardt (LM) algorithm was used as the learning algorithm. Zanetti et al. (2007) stated that the LM is better than the back propagation learning algorithm due to a faster convergence in the network learning. Detailed information about the LM can be seen in Mehdizadeh et al. (2016).

2.3 Gene Expression Programming

Gene expression programming (GEP) was first presented by Ferreira (2001). The GEP has been developed on the basis of Darwinian Evolution Theory. The GEP acts like the genetic algorithm (GA) and genetic programming (GP). The sample of individuals is used in the GEP. The individuals are selected on the basis of their merits. Genetic variations are exerted using genetic operators. The main difference among GA, GP and GEP is the nature of their individuals. In the GA, the individuals are linear strings of fixed length (chromosomes), while the individuals in the GP are non-linear entities of different sizes and shapes (parse trees). Nevertheless in the GEP, the individuals are encoded as linear strings of fixed-length (similar to the GA), and then, they are expressed in the form of non-linear entities of different sizes and shapes (similar to the GP).

In the GEP, each gene is encoded in the form of expression tree which represents a solution for the problem. All expression trees are connected to each other using a linking function in the case of multi-gene chromosomes. Readers can refer to Ferreira (2001) to see more information about the GEP.

In this study, GeneXpro Tools program was used. The estimation process of the monthly rainfall using the GEP is as follows:

The first stage is to determine a suitable fitness function. In the present research, root mean square error (RMSE) was used as the fitness function. Terminals and functions sets are selected in the second stage. In fact, terminals set are independent variables including different parameters which may affect the estimation of rainfall. In this study, one to five previous monthly rainfall in the form of lagged data were considered as the terminals set. Also, four arithmetic operators $(+, -, \times, \div)$ along with mathematical functions $(x^2, x^3, e^x, Lnx, \sqrt{x}, \sqrt[3]{x}, Sinx, Cosx, Atanx)$ were used as the functions set. The third stage is to select the structure of the chromosomes including head size and the number of genes and chromosomes. The head size, the number of genes and chromosomes were selected 8, 3 and 30, respectively. The selection of linking function is the fourth stage. In this study, an addition

function was employed to link the selected three genes (or expression trees). The final stage is to choose the genetic operators and their rates. In this process, a combination of all operators including three types of transposition, three types of recombination, as well as mutation and inversion were utilized.

The values of used parameters in the GEP are given in Table 3.

2.4 Developing the GEP and ANN Models

In this study, one to five previous monthly rainfall in the form of lagged data were used as inputs to develop the different GEP and ANN scenarios. These scenarios are named as below,

- GEP1, ANN1: P_{t-1}
- GEP2, ANN2: P_{t-1}, P_{t-2}
- GEP3, ANN3: P_{t-1}, P_{t-2}, P_{t-3}
- GEP4, ANN4: P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}
- GEP5, ANN5: P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_{t-5}

For example, GEP1 and ANN1 are scenarios in which the monthly rainfall of previous month was used to estimate the rainfall of current month.

2.5 Introducing New GEP-ARCH and ANN-ARCH Hybrid Models

One of the main problems in modeling of hydrological time series such as rainfall is to determine the random component of the time series. In this direction, entering a series of random data may not affect the estimated values. In this study, the random component of the monthly rainfall time series data is modeled using the ARCH model. In the most hydrological studies, the main attention is focused on the mean of data. However, little attention has been paid to the variance changes over the time. Regarding the progress of conducted studies on the field of risk and uncertainty in water resources engineering, it is essential to develop modeling techniques for considering the variance changes with respect to the time. The present study attempts to create novel methods by combining the ARCH time series model with the GEP and ANN to enhance the estimation accuracy of the rainfall time series.

The ARCH model which was first presented by Engle (1982) is a non-linear time series model for modeling the variance changes over the time as below,

Table 3 The values of parameters used for estimation of the monthly rainfall using GEP		
Number of genes		3
Number of chromosomes		30
Head size		8
Mutation rate		0.044
Inversion rate		0.1
Gene recombination rate		0.1
One-point recombination rate		0.3
Two-point recombination rate		0.3
Gene transposition rate		0.1
Insertion sequence transposition rate		0.1
Root insertion sequence transposition rate		0.1

$$\sigma_t^2 = a_0 + \sum_{i=1}^{m} b_i \varepsilon_{t-i}^2 \tag{1}$$

$$\varepsilon_t = \sigma_t z_t \tag{2}$$

where, σ_t^2 is the conditional variance, ε_t is the error phrase or model residual, a_0 and b_i are the model parameters, m is the order of model and $z_t \sim N(0, 1)$ is the normalized and standardized dataset.

Fitting steps of the ARCH model are as follows,

1- The best AR (Eq. 3) or ARMA (Eq. 4) is fitted on the stationary data series (z_t),

$$z_t(p) = \sum_{i=1}^{p} (\phi_i . z_{t-i}) + \varepsilon_t \tag{3}$$

$$z_t(p, q) = \sum_{i=1}^{p} (\phi_i . z_{t-i}) - \sum_{i=1}^{q} (\theta_i . \varepsilon_{t-i}) + \varepsilon_t \tag{4}$$

where, p is the order of AR model, q is the order of MA model, ϕ_i and θ_i are the coefficients of the models.

2- The error rates or residuals (ε_t) are calculated by Eqs. 3 or 4.
3- ε_t^2 is calculated using step 2,
4- ARCH model is fitted on the obtained ε_t^2 values and then Eqs. 1 and 2 are used to obtain the random component of rainfall data (ε_t).
5- The final step is the combination of ARCH time series model with the GEP and ANN models.

Since the rainfall time series consist of the deterministic and random parts, the results of the GEP and ANN models are combined with the results of the ARCH model as below,

$$P_t = D_t + R_t \tag{5}$$

where, D_t is the generated deterministic part of the rainfall time series using GEP and ANN models, R_t (or ε_t) is the obtained random part of the rainfall time series using ARCH model and P_t is the final value of the rainfall time series. Both D_t and R_t are the normalized and standardized data of the monthly rainfall time series.

Equations 6 (Delleur and Karamouz 1982) and 7 were used to normalize and standardize the monthly rainfall time series, respectively,

$$N_t = Ln(X_t + c) \tag{6}$$

$$Z_t = \frac{N_t - \mu}{\sigma} \tag{7}$$

where, N_t is the normalized values of the rainfall data, X_t is the actual values of the rainfall data, c is a constant, Z_t is the standardized values of the rainfall data. Also, μ and σ are the mean and standard deviation of the normalized monthly rainfall data, respectively, in monthly scale. It should be noted that the c coefficient in Eq. 6 is obtained using trial and error procedure to achieve the lowest value for the skewness coefficient.

Table 4 Calculated statistical indices for the different GEP and GEP-ARCH models in the testing stage

Scenarios	Rasht		Gorgan		Urmia		Kerman		Zahedan	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
GEP										
GEP1	75.31	0.330	31.43	0.197	23.78	0.284	13.17	0.319	12.15	0.223
GEP2	74.08	0.350	31.80	0.181	23.82	0.279	12.74	0.366	12.68	0.106
GEP3	73.29	0.363	31.41	0.199	24.07	0.268	14.13	0.249	11.95	0.233
GEP4	74.59	0.342	31.64	0.189	24.34	0.251	12.59	0.375	12.16	0.181
GEP5	75.88	0.322	31.65	0.188	24.61	0.237	12.76	0.359	12.19	0.194
GEP-ARCH										
GEP1-ARCH	57.14	0.941	15.66	0.963	17.64	0.932	5.93	0.887	8.79	0.572
GEP2-ARCH	56.61	0.951	16.31	0.956	16.20	0.947	5.95	0.867	**8.45**	**0.619**
GEP3-ARCH	64.86	0.924	18.99	0.944	16.19	0.937	7.44	0.783	8.59	0.602
GEP4-ARCH	**56.48**	**0.947**	**14.95**	**0.966**	**15.89**	**0.938**	**5.79**	**0.873**	8.53	0.607
GEP5-ARCH	57.46	0.939	16.59	0.962	16.24	0.935	6.21	0.857	9.10	0.549

2.6 Performance Evaluation Criteria

To investigate the performance of the GEP and ANN models, as well as the proposed GEP-ARCH and ANN-ARCH hybrid models in estimating the monthly rainfall, two statistical indicators including root mean square error (RMSE) and coefficient of determination (R^2) were used as below,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(E_i - O_i)^2}{n}} \tag{8}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}\left(E_i - \overline{E_i}\right)\left(O_i - \overline{O_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(E_i - \overline{E_i}\right)^2 \sum_{i=1}^{n}\left(O_i - \overline{O_i}\right)^2}} \right)^2 \tag{9}$$

Table 5 Calculated statistical indices for the different ANN and ANN-ARCH models in the testing stage

Scenarios	Rasht		Gorgan		Urmia		Kerman		Zahedan	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
ANN										
ANN1	74.52	0.346	32.10	0.172	24.62	0.240	13.07	0.329	12.43	0.128
ANN2	74.95	0.336	30.56	0.241	24.68	0.242	12.80	0.362	12.13	0.229
ANN3	75.84	0.342	31.17	0.210	24.70	0.240	13.07	0.328	12.21	0.188
ANN4	75.46	0.330	30.52	0.243	24.09	0.271	13.00	0.342	11.87	0.211
ANN5	77.24	0.301	30.68	0.235	23.95	0.289	12.77	0.362	11.65	0.236
ANN-ARCH										
ANN1-ARCH	52.94	0.949	**16.06**	**0.959**	14.80	0.953	6.13	0.864	8.62	0.609
ANN2-ARCH	57.61	0.948	17.43	0.957	**13.03**	**0.953**	6.12	0.859	8.76	0.573
ANN3-ARCH	53.80	0.918	19.24	0.941	13.09	0.945	5.67	0.893	8.60	0.592
ANN4-ARCH	**51.52**	**0.947**	17.53	0.957	14.01	0.952	5.88	0.873	8.48	0.629
ANN5-ARCH	55.15	0.938	19.10	0.940	14.74	0.932	**5.58**	**0.887**	**8.47**	**0.644**

Table 6 Results of the fitted ARCH model on the monthly rainfall time series during 1981–2012

| Station	Model order	Model coefficients		Fitted ARCH model
	m	$a0$	b_1	
Rasht	1	2.970	1.20×10^{-8}	$\sigma_t^2 = 2.970 + \left(1.20 \times 10^{-8} \times \varepsilon_{t-1}^2\right)$
Gorgan	1	1.793	5.49×10^{-8}	$\sigma_t^2 = 1.793 + \left(5.49 \times 10^{-8} \times \varepsilon_{t-1}^2\right)$
Urmia	1	2.369	7.17×10^{-8}	$\sigma_t^2 = 2.369 + \left(7.17 \times 10^{-8} \times \varepsilon_{t-1}^2\right)$
Kerman	1	0.618	2.37×10^{-8}	$\sigma_t^2 = 0.618 + \left(2.37 \times 10^{-8} \times \varepsilon_{t-1}^2\right)$
Zahedan	1	2.016	4.77×10^{-8}	$\sigma_t^2 = 2.016 + \left(4.77 \times 10^{-8} \times \varepsilon_{t-1}^2\right)$
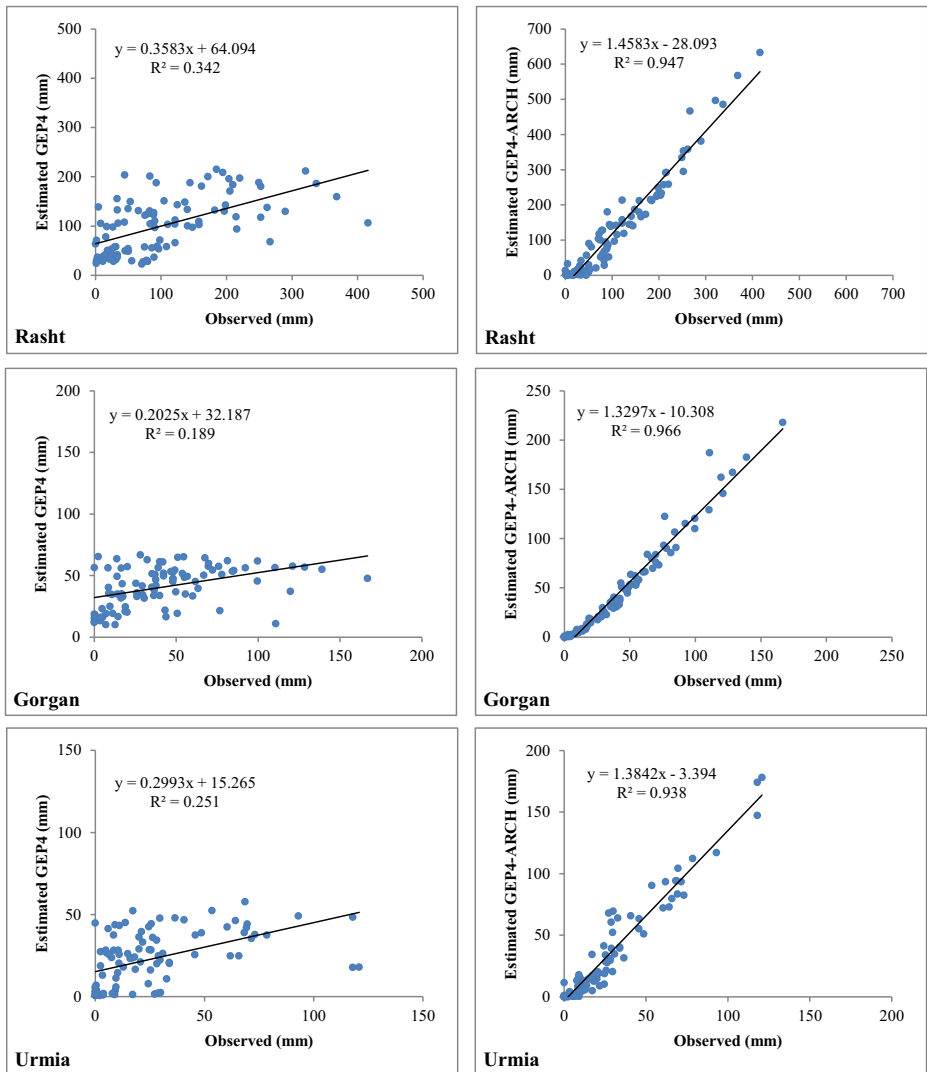
Fig. 3 Scatter plots between the observed and estimated rainfall for the best scenarios in the GEP-ARCH models and the corresponding GEP models during the testing stage

where, E_i is the ith estimated monthly rainfall using GEP, ANN, GEP-ARCH and ANN-ARCH models; O_i is the ith observed monthly rainfall; $\overline{E_i}$ is the average of the estimated monthly rainfall data; $\overline{O_i}$ is the average of the observed monthly rainfall data and n is the number of observations. RMSE is measured in mm in the present study.

3 Results and Discussion

In the present study, one to five previous monthly rainfall in the form of lagged data were used to estimate the monthly rainfall of current month. At first, the accuracy of GEP and ANN models was investigated at various climatic regions in Iran. Then, the performance of proposed novel models including GEP-ARCH and ANN-ARCH was evaluated and then compared to the GEP and ANN models.

3.1 Results of the GEP and ANN Models

The results of the GEP and ANN models including RMSE and R^2 statistical indices in estimating monthly rainfall during the testing stage are presented in Tables 4 and 5, respectively. As seen, the highest and lowest values of the RMSE indicator are observed in Rasht and Zahedan stations, respectively. In Rasht station with a humid climate, the GEP models in all
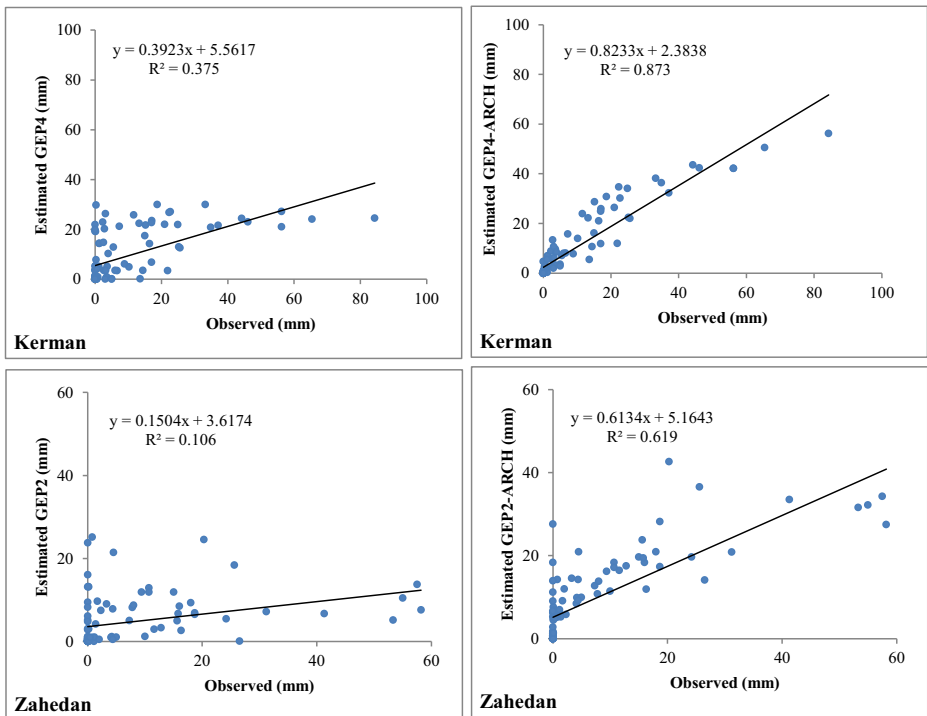
Fig. 3 (continued)

scenarios have higher precision than the corresponding ANN scenarios, except GEP1. In other stations, constant state is not observed, i.e. in some scenarios, the GEP performs better than the ANN and vice versa. The GEP is better than the ANN in three scenarios at Urmia and Kerman stations. However, the ANN had slightly better performance in four and three scenarios at Gorgan and Zahedan stations, respectively. As seen in Tables 4 and 5, the performance of scenarios are not necessarily improved by increasing the input variables in the GEP. For example, this condition is seen in Rasht and Urmia stations. In the mentioned stations, the GEP5 which uses full inputs has the lowest accuracy. In the ANN, the situation is different. The ANN5 in Urmia (RMSE = 23.95 mm), Kerman (RMSE = 12.77 mm) and Zahedan (RMSE = 11.65 mm) presents better performance than other scenarios with fewer inputs. Also,
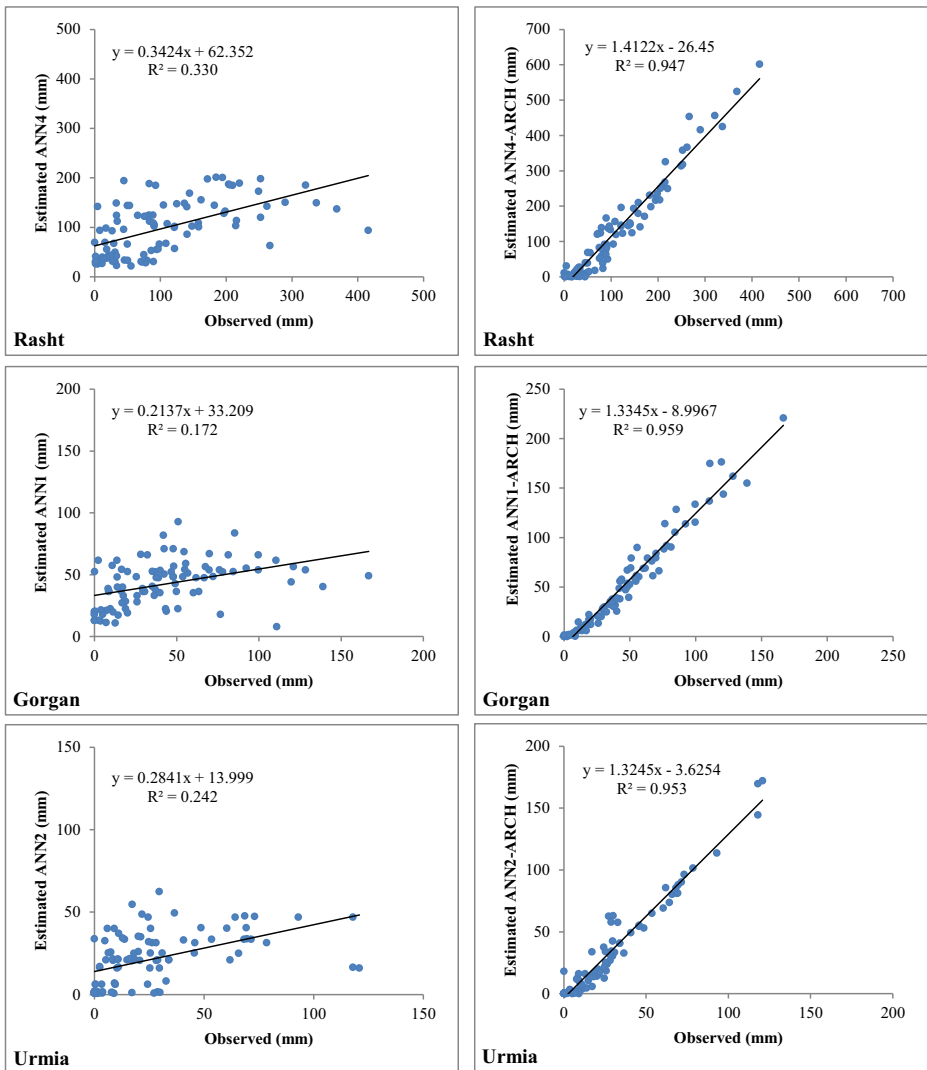
Fig. 4 Scatter plots between the observed and estimated rainfall for the best scenarios in the ANN-ARCH models and the corresponding ANN models during the testing stage

the ANN5 is the third superior model after ANN4 and ANN2, in Gorgan. The only exception in this field is Rasht station in which the ANN5 (similar to the GEP5) is the worst scenario compared to the others. The reason of this fact can be explained with attention to the type of climate of the mentioned station and the large differences between the values of the rainfall during various months.

3.2 Comparison of Novel GEP-ARCH and ANN-ARCH Hybrid Models with GEP and ANN

As mentioned before, the time series of hydrologic processes such as rainfall compose two parts; deterministic and random components. In this study, the deterministic part of the monthly rainfall time series was obtained by the GEP and ANN models. Moreover, the random component of rainfall time series was obtained using the ARCH model. Then, the obtained random component was combined and added with the results of the GEP and ANN models (according to Eq. 5). Results of the fitted ARCH model on time series of the monthly rainfall data are listed in Table 6. Also, calculated statistical indices for the proposed GEP-ARCH and ANN-ARCH hybrid models are given in the second part of the Tables 4 and 5, respectively. As seen, by combining the GEP and ANN with ARCH model, the performance of models is considerably improved for all of the stations and used scenarios in comparison with the GEP and ANN models. For example, the values of RMSE and R^2 statistical indices in
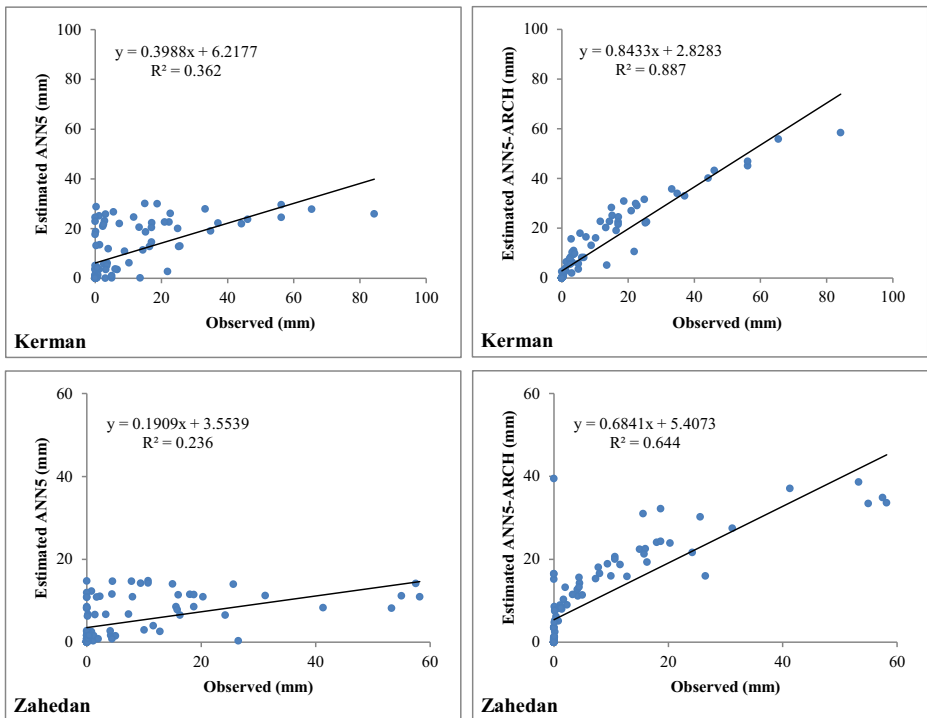

Fig. 4 (continued)

the GEP4 for Rasht station are 74.59 mm and 0.342, respectively. The mentioned statistics improve to 56.48 mm and 0.947, respectively in the GEP4-ARCH model. Similar trends were observed in other GEP-ARCH, as well as ANN-ARCH models for other stations. The performance comparison of the GEP-ARCH and ANN-ARCH models reveals that the GEP-ARCH scenarios are superior to the ANN-ARCH models in Gorgan. On the other hand, the ANN-ARCH scenarios in Rasht and Urmia stations present better results than the corresponding GEP-ARCH models (except for ANN2-ARCH in Rasht). Furthermore, the precision of the GEP-ARCH and ANN-ARCH models are slightly similar for the different scenarios in Kerman and Zahedan stations. The only exception is ANN3-ARCH in Kerman where the mentioned scenario has better performance than the GEP3-ARCH.
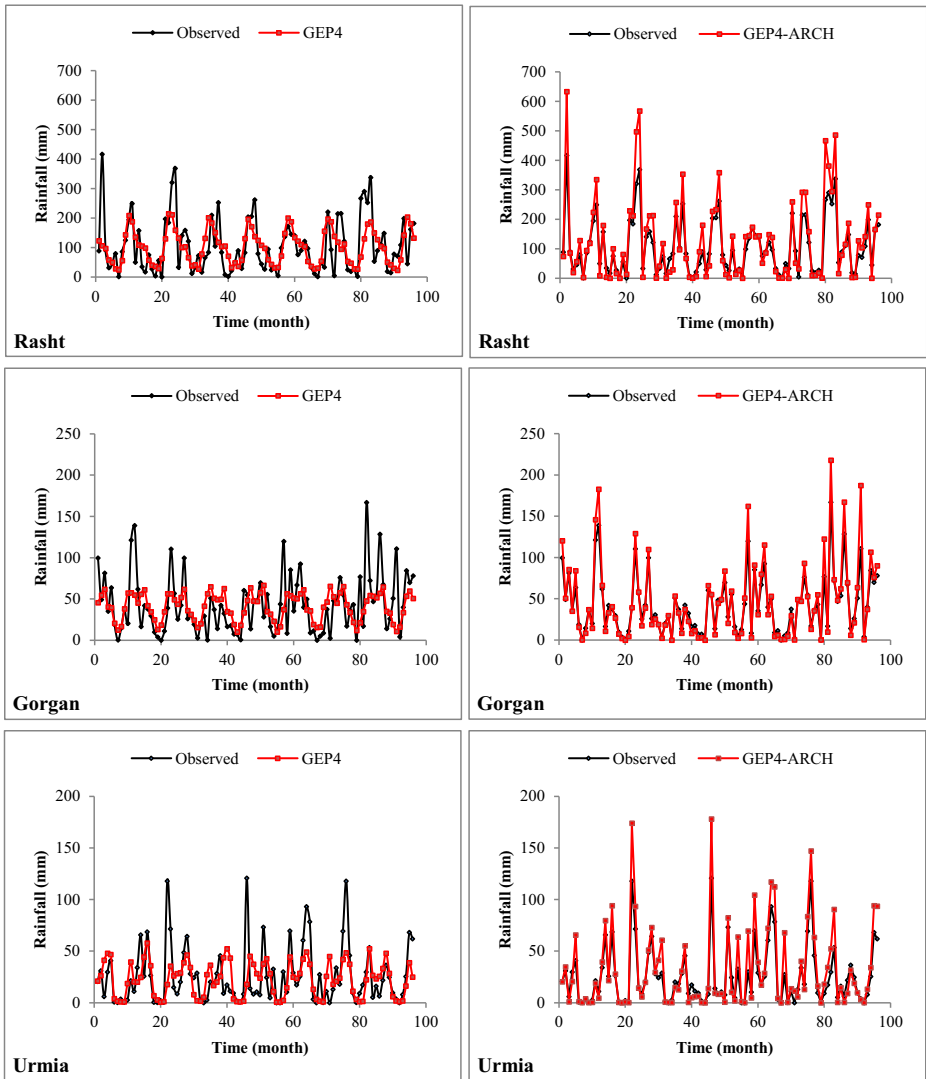
Fig. 5 Time series of the observed and estimated rainfall for the best scenarios in the GEP-ARCH models and the corresponding GEP models during the testing stage

To graphically illustrate the performance of proposed hybrid models compared to the GEP and ANN models, superior scenarios of the testing stage (with bolded statistical indices in Tables 4 and 5) in the hybrid models, as well as the corresponding scenarios in the GEP and ANN models were chosen. Scatter plots between the observed and estimated monthly rainfall for the best scenarios in the proposed GEP-ARCH and ANN-ARCH hybrid models, as well as the corresponding GEP and ANN models in the testing period are depicted in Figs. 3 and 4, respectively. It is clear from the scatter plots that the estimated monthly rainfall data by the proposed GEP-ARCH and ANN-ARCH hybrid models are closer to the observed data than those obtained by the GEP and ANN models. The highest correlations are observed in Gorgan, Rasht and Urmia stations (see also other scenarios in Tables 4 and 5). On the other hand, the lowest correlation is seen for Zahedan as a hyper-arid region. By assuming $y = ax + b$ equation (y is the estimated rainfall and x is the observed rainfall) for fitted lines to the models, it is clear that a (slope) and b (intercept) coefficients in the GEP-ARCH and ANN-ARCH hybrid models are closer to 1 and 0, respectively in comparison with the GEP and ANN models.

Figures 5 and 6 demonstrate the time series of observed and estimated monthly rainfall for the superior scenarios in the hybrid models and the corresponding GEP and ANN scenarios during the testing period. It is clear from the Figures that a fixed trend is observed in the time series of estimated monthly rainfall data for all stations in the GEP and ANN models. In addition, the mentioned models are not able to estimate peak points of the monthly rainfall data. However, the proposed GEP-ARCH and ANN-ARCH models perform better than the
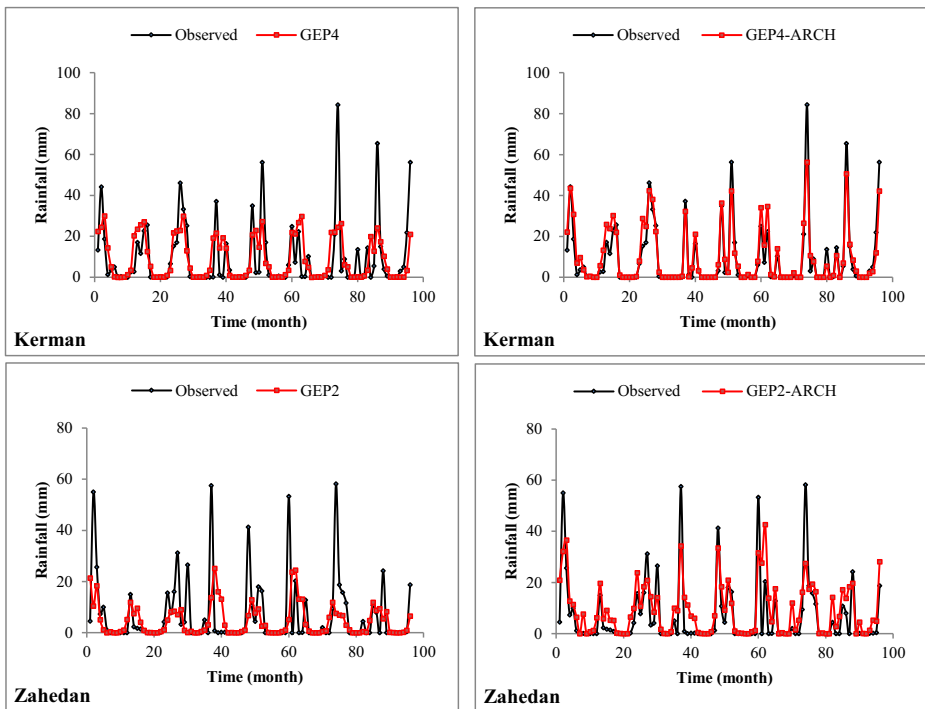

Fig. 5 (continued)

GEP and ANN in estimating time series of the monthly rainfall, as well as peak values of the rainfall data. It is seen from the Figs. 5 and 6 that both proposed models overestimate peak values of the rainfall time series in Rasht, Gorgan and Urmia stations. Furthermore, the peak values of rainfall data are underestimated in Kerman and Zahedan stations as arid and hyper-arid climates, respectively. Similar to the obtained results, Kisi and Cimen (2012), Venkata Ramana et al. (2013), Feng et al. (2015) and Shenify et al. (2016) reported that peak points of the rainfall data were closely matched to the observed data using the hybrid models such as wavelet-support vector machine and wavelet-neural networks.
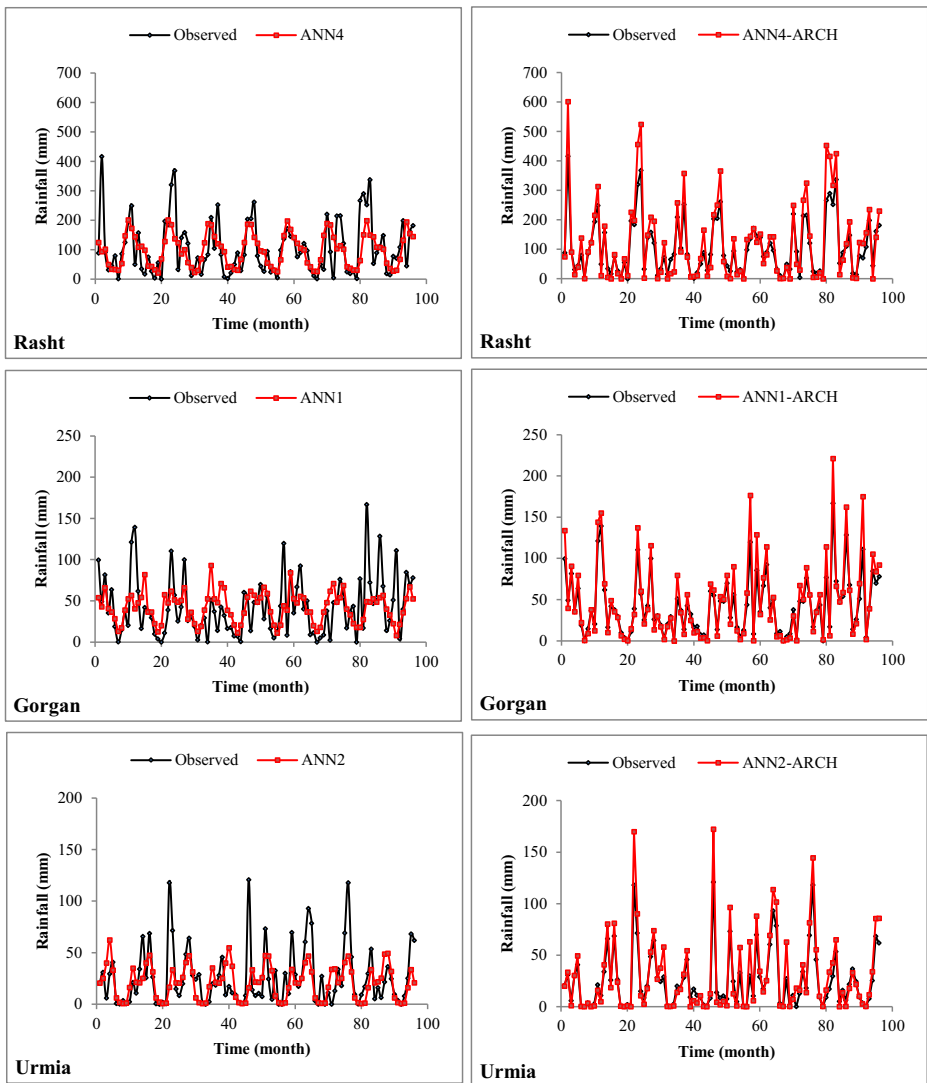

Fig. 6 Time series of the observed and estimated rainfall for the best scenarios in the ANN-ARCH models and the corresponding ANN models during the testing stage
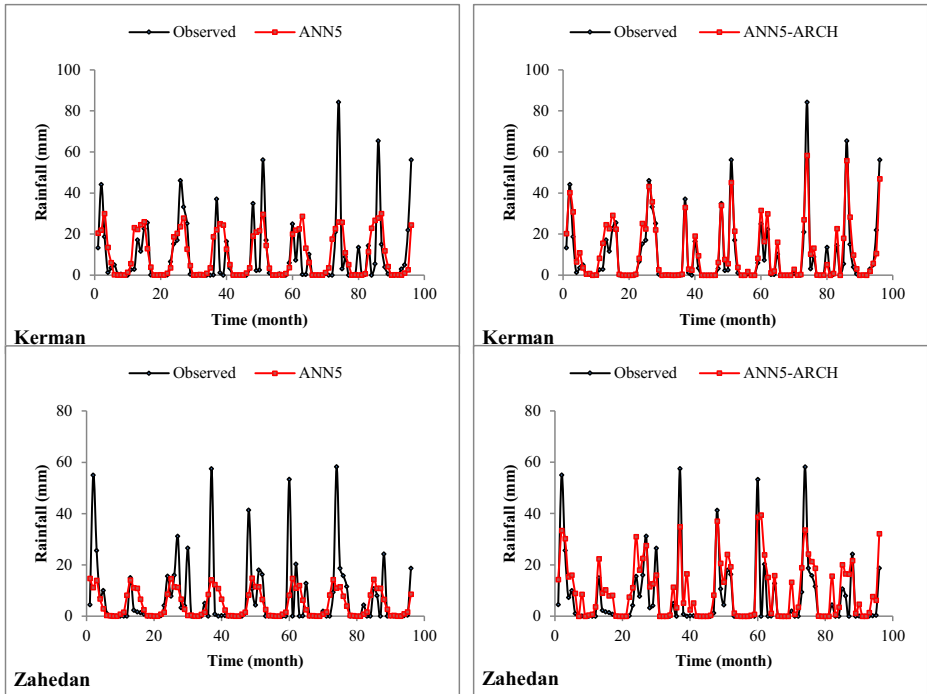
Fig. 6 (continued)

4 Conclusion

In the present study, at first, the capability of the GEP and ANN in estimating monthly rainfall time series was investigated at five stations with various climatic conditions in Iran. The results showed that both models presented slightly similar performances. Moreover, the models' accuracies were not acceptable. In the second part, two novel models were introduced by combining the GEP and ANN intelligent models with the ARCH time series model. So, the GEP-ARCH and ANN-ARCH hybrid models were developed. The obtained results demonstrated that the proposed hybrid models yielded better performances compared to the GEP and ANN models. On the other hand, these hybrid models estimated the same trends of the observed rainfall data, as well as peak values of the rainfall time series with high accuracy compared to the GEP and ANN models. However, the proposed hybrid models are slightly over estimating the peak points of rainfall data in Rasht (Humid), Gorgan (sub-humid) and Urmia (semi-arid) stations. Moreover, the mentioned models had underestimation at selected stations in arid and hyper-arid climates. It is expected that it can be achieved better results by combining the intelligent models with time series models such as ARCH model in the hydrological time series modeling. Moreover, it is recommended that other types of ARCH model such as GARCH, EGARCH and Non-linear GARCH to be used in the next studied to model the random part of hydrological time series (e.g. rainfall). Further studies can also be accomplished to present other novel models in order to more accurate modeling of the rainfall time series, especially peak points of the rainfall data.

References

Abbot J, Marohasy J (2014) Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmos Res 138:166–178

Ashraf B, Yazdani R, Mousavi-Baigy M, Bannayan M (2014) Investigation of temporal and spatial climate variability and aridity of Iran. Theor Appl Climatol 118(1–2):35–46

Behmanesh J, Mehdizadeh S (2017) Estimation of soil temperature using gene expression programming and artificial neural networks in a semiarid region. Environ Earth Sci. https://doi.org/10.1007/s12665-017-6395-1

Chinchorkar SS, Patel GR, Sayyad FG (2012) Development of monsoon model for long range forecast rainfall explored for Anand (Gujarat-India). Int J Water Resour Environ Eng 4(11):322–326

Delleur JW, Karamouz M (1982) Uncertainty in reservoir operation. Optimal Allocation of Water Resources (Proceedings of the Fxeter Symposium), IAHS Publication no. 135:7–16

Ebtehaj I, Bonakdari H, Zaji AH, Azimi H, Sharifi A (2015) Gene expression programming to predict the discharge coefficient in rectangular side weirs. Appl Soft Comput 35:618–628

Engle RF (1982) Autoregressive conditional heteoscedasticity with estimates of the variance of United Kingdom inflations. Econometrica 50(4):987–1007

Feng Q, Wen X, Li J (2015) Wavelet analysis-support vector machine coupled models for monthly rainfall forecasting in arid regions. Water Resour Manag 29(4):1049–1065

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

Gocic M, Motamedi S, Shamshirband S, Petkovic D, Ch S, Hashim R, Arif M (2015) Soft computing approaches for forecasting reference evapotranspiration. Comput Electron Agric 113:164–173

Haykin S (1998) Neural networks-a comprehensive foundation, 2nd edn. Prentice-Hall, Upper Saddle River, pp 26–32

He X, Guan H, Qin J (2015) A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall. J Hydrol 527:88–100

Imani M, You RJ, Kuo CY (2014) Forecasting Caspian Sea level changes using satellite altimetry data (June 1992–December 2013) based on evolutionary support vector regression algorithms and gene expression programming. Glob Planet Chang 121:53–63

Kashid SS, Maity R (2012) Prediction of monthly rainfall on homogeneous monsoon regions of India based on large scale circulation patterns using genetic programming. J Hydrol 454–455:26–41

Khalili K, Nazeri Tahroudi M, Mirabbasi R, Ahmadi F (2016) Investigation of spatial and temporal variability of precipitation in Iran over the last half century. Stoch Environ Res Risk Assess 30(4):1205–1221

Kisi O, Cimen M (2012) Precipitation forecasting by using wavelet-support vector machine conjunction model. Eng Appl Artif Intell 25(4):783–792

Marti P, Shiri J, Duran-Ros M, Arbat G, de Cartagena FR, Puig-Bargues J (2013) Artificial neural networks vs. Gene Expression Programming for estimating outlet dissolved oxygen in micro-irrigation sand filters fed with effluents. Comput Electron Agric 99:176–185

Mehdizadeh S, Behmanesh J, Khalili K (2016) Comparison of artificial intelligence methods and empirical equations to estimate daily solar radiation. J Atmos Sol Terr Phys 146:215–227

Mehdizadeh S, Behmanesh J, Khalili K (2017) Application of gene expression programming to predict daily dew point temperature. Appl Therm Eng 112:1097–1107

Mekanik F, Imteaz MA, Gato-Trinidad S, Elmahdi A (2013) Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. J Hydrol 503:11–21

Moustris KP, Larissi IK, Nastos PT, Paliatsos AG (2011) Precipitation forecast using artificial neural networks in specific regions of Greece. Water Resour Manag 25(8):1979–1993

Partal T, Kisi O (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. J Hydrol 342(1–2):199–212

Ramirez MCV, de Campos Velho HF, Ferreira NJ (2005) Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. J Hydrol 301(1–4):146–162

Shenify M, Danesh AS, Gocic M, Taher RS, Abdul Wahab AW, Ghani A, Shamshirband S, Petkovic D (2016) Precipitation estimation using support vector machine with discrete wavelet transform. Water Resour Manag 30(2):641–652

Shiri J, Keshavarzi A, Kisi O, Iturraran-Viveros U, Bagherzadeh A, Mousavi R, Karimi S (2017) Modeling soil cation exchange capacity using soil parameters: assessing the heuristic models. Comput Electron Agric 135: 242–251

UNEP (1992) World atlas of desertification. The united nations environment programme (UNEP), London

Venkata Ramana R, Krishna B, Kumar SR, Pandey NG (2013) Monthly rainfall prediction using wavelet neural network analysis. Water Resour Manag 27(10):3697–3711

Wu CL, Chau KW, Fan C (2010) Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. J Hydrol 389(1–2):146–167

Yassin MA, Alazba AA, Mattar MA (2016) A new predictive model for furrow irrigation infiltration using gene expression programming. Comput Electron Agric 122:168–175

Zanetti SS, Sousa EF, Oliveira VP, Almeida FT, Bernardo S (2007) Estimating evapotranspiration using artificial neural network and minimum climatological data. J Irrig Drain Eng 133(2):83–89