International Journal of Environmental Science and Technology (2022) 19:11919–11934
https://doi.org/10.1007/s13762-022-04395-2

ORIGINAL PAPER

Coupled data pre-processing approach with data intelligence models for monthly precipitation forecasting

M. R. Nikpour[1] · S. Abdollahi[2] · H. Sanikhani[3] · J. Raeisi[4] · Z. M. Yaseen[5,6,7]

Received: 21 February 2021 / Revised: 9 February 2022 / Accepted: 11 July 2022 / Published online: 29 July 2022

Abstract

For effective and reliable management of water resources, the accurate forecasting of precipitation patterns is highly essential. The precipitation process is a complex hydrological component that is influenced by several hydro-meteorological variables, and it is varied from one region to another. In this study, the performance of coupled and standalone data intelligence models including wavelet artificial neural network (WANN), wavelet gene expressing programming (WGEP), artificial neural networks (ANN), and gene expressing programming (GEP) is undertaken for precipitation forecasting at four stations located in Iran (i.e., Ardabil, Khalkhal, Meshginshahr, and Parsabad). In this regard, monthly precipitation data are utilized from 1997 to 2016. The developed forecasting models are constructed using correlated lag time information. Two statistical performance metrics: coefficient of determination (R^2) and root mean square errors (RMSE), are calculated for forecasting accuracy inspection. The obtained results indicated the capacity of the WANN coupled model as a superior forecasting model compared with the other data intelligence models. This was observed at the four investigated meteorological stations. However, the potential of the discrete wavelet transform has also demonstrated an enhancement in the predictability performance of the GEP model. This is indicating the capability of the pre-processing data time series as a prior stage for the forecasting process. Further, the results evidence the capability of the WANN coupled model in capturing the peak values of the precipitation. Overall, the applied coupled data intelligence model provided a significant contribution to the precipitation forecasting at this region contributing to the base knowledge of water resources engineering.

Keywords Precipitation · Artificial neural network · Genetic expressing programming · Wavelet transforms

Editorial responsibility: Samareh Mirkia.

✉ M. R. Nikpour
m_nikpour@uma.ac.ir

[1] Department of Water Engineering, Faculty of Agriculture and Natural Resources, University of Mohaghegh Ardabili, Ardabil, Iran

[2] Regional Water Company of Charmahal and Bakhtiari, Water Resources Management Company, Ministry of Energy, Shahrekord, Iran

[3] Department of Water Sciences and Engineering, Faculty of Agriculture, University of Kurdistan, Sanandaj, Iran

[4] Water and Wastewater Company of Charmahal and Bakhtiari, National Water and Wastewater Engineering Company, Ministry of Energy, Shahrekord, Iran

[5] Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia

[6] USQ's Advanced Data Analytics Research Group, School of Mathematics Physics and Computing, University of Southern Queensland, Darling Heights, QLD 4350, Australia

[7] New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar 64001, Iraq

Introduction

Research background

Precipitation is a significant variable in the hydrologic cycle and is different in time and space. Therefore, monthly and seasonal forecasting of precipitation gives an important information about studying runoff, groundwater, sediment, flood, and all of the phenomena related to water resources management (Everingham et al. 2008; Wu et al. 2010). Precipitation can be forecasted using two categories of approaches including dynamical and empirical methods. Dynamical methods are based on the rules of physics used for forecasting climate (Lim et al. 2009; DelSole and Shukla 2012; Schepen et al. 2012). The logic behind empirical methods is determining the relevant characteristics of previous records of variables used for forecasting and applying them to forecast in the future. However, on the other hand, the empirical models showed a reliable alternative for modeling hydrological processes. The most commonly used empirical methods include autoregressive model (AR), autoregressive moving-average models (ARMA), disaggregation multivariate model, artificial neural networks (ANN), genetic programming (GP), genetic expressing programming (GEP), neuro-fuzzy, and support vector machine (SVM) (Abbot and Marohasy 2012; Mekanik et al. 2013; Deo and Şahin 2015; Mehdizadeh et al. 2018; Danandeh Mehr and Nourani 2018; Zeynoddin et al. 2018).

Research significant

It should be noted that the introduced linear and nonlinear methods have limitations with the non-linearity and non-stationary data features (Ali et al. 2020; Salih et al. 2020). Precipitation pattern is highly stochastic natural problem that is influenced by several climate parameters such air temperature, humidity, wind, and the amount of evaporation from the earth surface (Dawkins et al. 2022). Matter of fact, hydrological data time series modeling is required data cleaning, clustering, and preprocessing. Hence, if the input data are not pre-processed with the appropriate technique, most models cannot deal with the non-stationary aspect of time series. The methods for handling non-stationary data are less advanced than those for stationary data. Among several data pre-processing approaches, wavelet transform is one of the distinguished techniques (Nourani et al. 2009a). Wavelet transform is one of the most effective methods for non-stationary data; this method has been studied in many fields outside of water resources engineering and hydrology (Adamowski and Sun 2010). One of the main features of wavelet transform is giving applicable decompositions of main time series. Also, the data obtained by wavelet analysis can enhance the ability of a forecasting model by capture applicable information on different resolution levels (Abdollahi et al. 2017).

Literature review and research motivation

Wavelet transform has been used in different situations and can be applied as a preprocessing and hybrid technique to improve some hydrological models such as precipitation, streamflow, rainfall–runoff, groundwater, snowmelt, drought, regionalization of watersheds, etc. (Lane 2007; Schaefli et al. 2007; Adamowski 2008a, b; Chou 2011; Dökmen and Aslan 2013; Miao et al. 2014; Feng et al. 2015; Christodoulou et al. 2017; Rathinasamy et al. 2017; Ahani et al. 2018, 2020a, b, c; Rezaei et al. 2021). A combination of mentioned models with wavelet can be used for forecasting precipitation time series (Partal and Kişi 2007; Kisi and Cimen 2012; Komasi and Sharghi 2016; Roushangar et al. 2018); especially, some researchers have used the combination of WT with ANN for forecasting the precipitation (Mwale and Gan 2005; Mwale et al. 2007; Nourani et al. 2009a; Partal and Cigizoglu 2009; Kuo et al. 2010; Chou 2011; Venkata Ramana et al. 2013; Miao et al. 2014; He et al. 2015; Shafaei et al. 2016; Arab Amiri et al. 2016; Zhang et al. 2018). Also, some researchers have coupled genetic programming-based models with WT (Kisi and Shiri 2011; Dabhi and Chaudhary 2014; Shoaib et al. 2015; Sezen and Partal 2017; Abdollahi et al. 2017). In these researches, it was found that the accuracy of hybrid models trained or calibrated with decomposed data is higher than the common models with row time series inputs. In this regard, Ghamariadyan and Imteaz (2021) developed a WANN model for medium-term rainfall prediction for 1, 3, 6, and 12 months in Queensland (Australia). The results revealed that the WANN improved the average prediction accuracy in terms of RMSE compared to the ANN. In another study, Wu et al. (2021) were utilized the wavelet transformation (WT), long short-term memory (LSTM) and autoregressive integrated moving average (ARIMA) methods to forecast the monthly precipitation during 1967–2017 at three meteorological stations in the Northeast China. They also developed a new hybrid model wavelet-ARIMA-LSTM (W-AL) of monthly precipitation time series. They found that the W-AL had higher prediction accuracy in monthly precipitation predicting than ARIMA and LSTM. Falayi et al. (2022) applied the wavelet transformation analysis and nonlinear dynamic time series techniques for investigating the chaotic behavior of monthly rainfall data. They observed the occurrences of strong oscillations in the rainfall at the West Africa stations. Wang et al. (2022) proposed an innovative application of the wavelet decomposition–prediction–reconstruction model (WDPRM) to enhance the accuracy of medium and long-term precipitation prediction. The proposed model was

successfully verified by using precipitation values from the Wujiang River Basin for 1961–2018.

Although several studies have been conducted on this aspect, limited research has been established on modeling precipitation using those coupled models. In addition, precipitation modeling is varied from one region to another due to the hydrological characteristics of that particular region. Furthermore, those models have based machine learning models that behave differently from one case to another. Hence, the exploration of the feasibility of those models is still ongoing era of modeling research.

Research objectives

In this study, two types of data intelligence models (coupled models of WANN and WGEP, standalone models of ANN and GEP) are developed for monthly precipitation time series forecasting at four stations including Ardabil, Khalkhal, Meshginshahr, and Parsabad. A data span covering the period of (1997–2016) is used to build the forecasting models. The main idea of investigating four different meteorological stations is to understand the generalization capacity of the

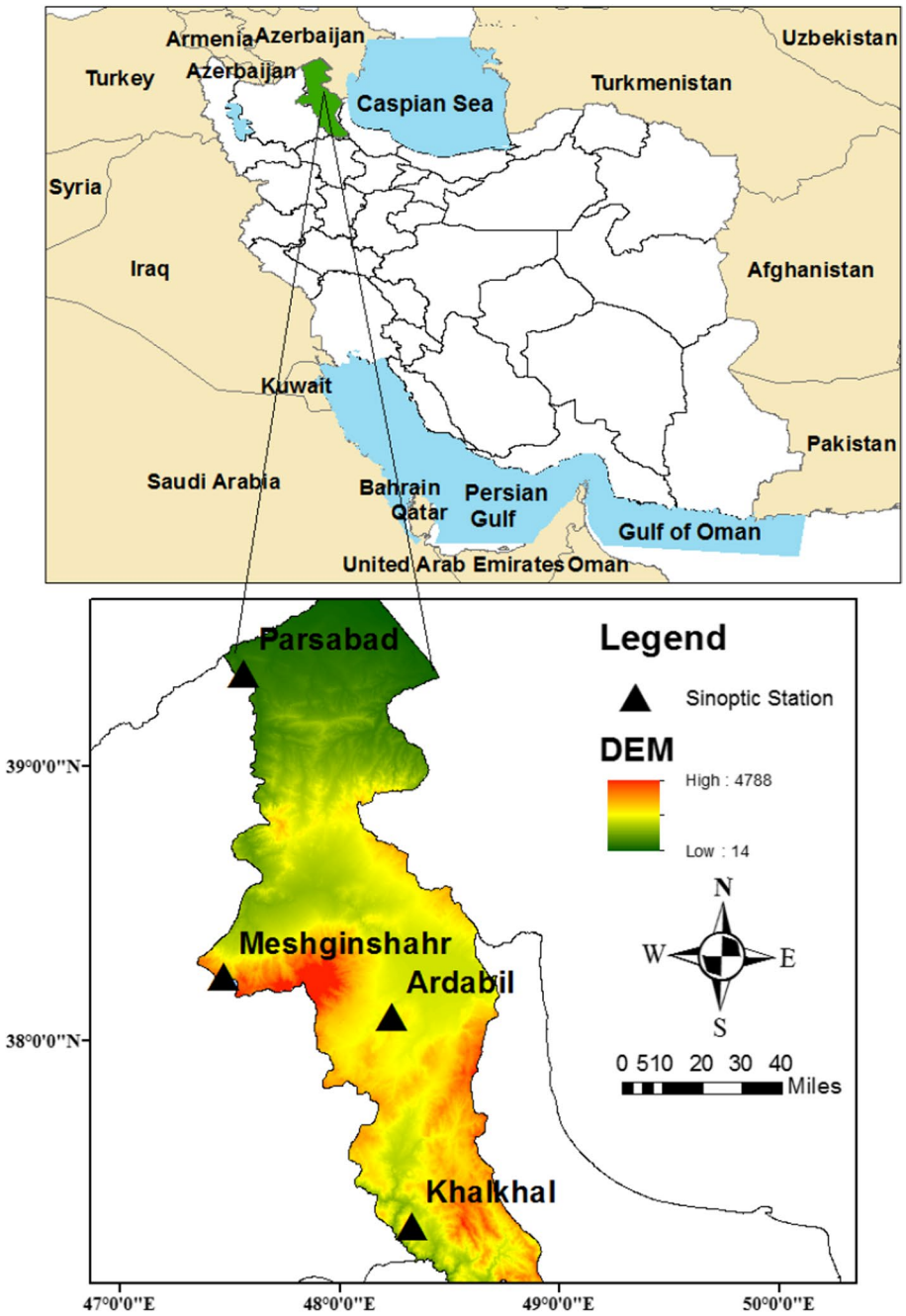

Fig. 1 Location of the Ardabil Province and study stations

Table 1 Geographical coordinate and statistical summary of some main climate variables

Parameter	Station			
	AL	KL	MR	PD
Latitude (E)	38° 8′	39° 21′	38° 14′	37° 10′
Longitude (N)	48° 11′	47° 27′	47° 25′	48° 19′
Elevation(m)	1335.2	72.6	1561.0	1796.0
Mean precipitation	291	275	388.3	368
Mean temperature	9.4	15.3	11	8.5
Mean humidity	71%	71%	58%	64%

proposed forecasting models to capture different patterns. Based on the attained predictability performance, a comprehensive assessment and evaluation are discussed.

Materials and methods

Study area and data description

The Ardabil Province in North-western Iran, located between latitude 37° 45′ to 39° 42′ N and longitude 47° 03′ to 48° 55′ E (Fig. 1), was investigated in this research. The geographical information and the mean observed climate data including mean precipitation, mean temperature, mean humidity for four synoptic stations, Ardabil (AL), Khalkhal (KL), Meshginshahr (MR), and Parsabad (PD), between 1997 and 2016, are presented in Table 1. In this study, before the application of precipitation data, a homogeneity test was used for the time series. The results indicated that the time series are almost homogeneous for the next analysis steps. Additionally, the missing data were filled using a simple regression method and more correlated neighbor stations.

Wavelet transform

Wavelet transform has been applied in different sciences, in which no stationary time series have been used, such as earthquake engineering, geophysics, hydrology, and climatology. The concept of wavelet transform can be understood by comparing it with the Fourier transform. Although the frequency content of the signal can be obtained by Fourier transform, this transform can't describe the localization of the time–frequency signal. This problem can be solved by wavelet transform. When a signal is converted by a wavelet transform, it is taken to 3-D space, containing, time, scale, and quantity. Therefore, wavelet transform can localize signal (Lafrenière and Sharp 2003). Wavelet transform consists of two forms: continuous wavelet transform called CWT and

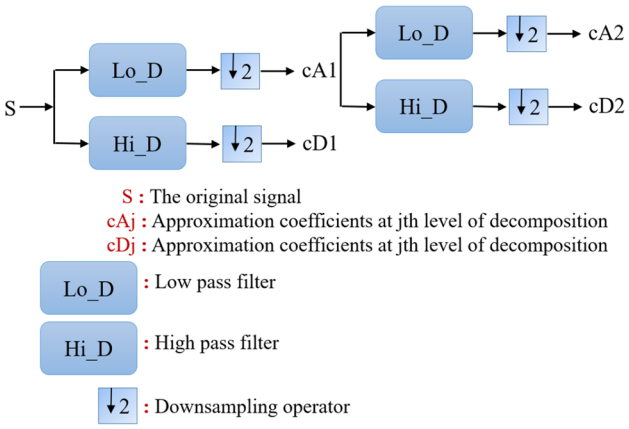

Fig. 2 Schematic representation of multiresolution analysis

discrete wavelet transform, namely DWT. The continuous wavelet transform is defined as below:

$$W(s,b) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t) \cdot \psi^* \left(\frac{t-b}{s} \right) \mathrm{d}t \tag{1}$$

where b indicates the translation of the wavelet over time, s represents the scale factor that determining the wavelet bandwidth ($s > 0$), $W(s,b)$ represents the wavelet coefficient determined by the application of transform on time series in scale s and time delay b, $\psi(.)$ is the mother wavelet or transformation function, $*$ is the complex conjugate (Cannas et al. 2006).

A signal can be decomposed by continuous wavelet transform. For this purpose, CWT uses many continuously varying scale and translation parameters. It needs considerable computing power when wavelet coefficients are computed for all possible scales. The run of DWT is simpler than CWT, and its calculations are less time-consuming in comparison with CWT. The values of scale and translation parameters of DWT are discretized as (Adamowski 2008a):

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{\left|S_0^j\right|}} \Psi \left(\frac{t - k b_0 s_0^j}{s_0^j} \right) \tag{2}$$

where j and k are integers, S_0 represents specified fined dilation step, and $S_0 > 1$, b_0 represents the location parameter, $\psi(.)$ is the function of the mother wavelet. In this study, the Mallat (1989) filtering process is applied for decomposing the time series. For this purpose, the time series is passed through the low-pass and high-pass filters in the same way. The outcomes obtained by the low-pass filters represent the high-scale and low-frequency wavelets, named approximation wavelets; they indicate the gross and slow-changing features of the signal. Also, the outcomes obtained by high-pass

filters are low-scale and high-frequency wavelets, namely detail wavelets; the rapidly changing features of the signal can be measured by detail coefficients. In the next decomposition level, approximation and detail wavelets can be determined by decomposing the approximation sub-time series. In Fig. 2, a two-level wavelet decomposition using DWT is shown. After adding all the detail sub-time series and the approximation sub-time series of the last decomposition level, the main signal can be reconstructed.

Gene expression programming

For the first time, Ferreira (2006) developed GEP model according to an evolutionary method. The GEP structure is a combination of GP and genetic algorithm (GA). On the other hand, the GEP method is the result of combining linear chromosomes with constant lengths like the GA and ramified architecture of different sizes and shapes as GP. In the GEP model, the initial population by the random generation of chromosomes is first made. Then, initial population is used for generating the chromosomes for the beginning of GEP modeling. In the following, expression trees can be used for expressing produced chromosomes. After evaluating the effectiveness of each chromosome, the best ones are chosen to modify and reconstruct new chromosomes. Genetic operators such as mutation, inversion, inversion sequence (IS) transposition, root insertion sequence (RIS), single or double crossover, gene crossover, and gene transposition mutation are used for modifying and reconstructing chromosomes (Ferreira 2006). In this study, GeneXproTools5 software has been used for developing and running models based on GEP. Five stages for forecasting the GEP process have been presented as follows: (1) choose terminal sets, independent variables; (2) select function sets, four main operators, accompanying with mathematical functions; (3) measure the accuracy of the model based on an index; (4) choose control components, numeral components, and quality variables values; (5) obtain results by choosing the condition for stopping the running program.

The precipitation (P) is modeled based on five stages using a GEP method. Initially, the fitness function should be selected. The performance of the obtained program is assessed based on the fitting function. This performance is evaluated using a relative error index. Also, the fitness function is used to select limitations for operating and accuracy. For this purpose, it enables accurate coordinating from achieved solutions with needful accuracy (Ferreira 2006). The fitness function (f_i) is determined as below.

$$f_i = \sum_{j=1}^{n} \left(R - \left| \frac{P_{(ij)} - T_j}{T_j} \times 100 \right| \right) \tag{3}$$

where R, $P_{(ij),}$ and T_j indicate selected limitations, the value forecasted by the individual program i for fitness case j and target value corresponding to fitness case j, respectively. In the related literature, the term presented inside modulus is called relative error and it refers to accuracy. If $\left|P_{(ij)} - T_j\right| \leq 0.01$, This condition is called entire fitness, and the value of accuracy will be equal to zero and $f_i = f_{\max} = nR$, consequently. Root mean square error (RMSE) index is used for obtaining the value of error in the fitness function. In the next step, the chromosomes will be generated by choosing the terminal and function set. The value of the outcome variable is determined as below.

$$Y = f(X^m) \tag{4}$$

where X^m represents the m dimensional input containing the variables of $x_1, x_2, \ldots, x_m$ and Y indicates the output variable. In this research, X_i represents previous precipitation with delay time and Y represents precipitation value in next times. In this study, selected terminals set include $P_t = \{P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_{t-5}, P_{t-6}, P_{t-7}, P_{t-8}, P_{t-9}, P_{t-10}, P_{t-11}, P_{t-12}\}$, in which $P_{(t)}$ represents precipitation in time t. The choosing of function set should include all the necessary function to have an appropriate guess for this function. In this study, four mathematical operators (+, −, ×, ÷) were utilized as functions set. In the next stage, chromosomal architecture should be determined. In this study, head length and number of genes have been selected 8 ($h = 8$) and 3 genes per chromosome, respectively. It should be noted that 30 chromosomes have been selected in each run. In the next step, the linkage function should be selected to defining the relationship between sub-trees. In this research, three considered genes coupled using total function. In the following, it is necessary determining genetic operators and corresponding rates. In this research, an inversion and a mutation probabilistic

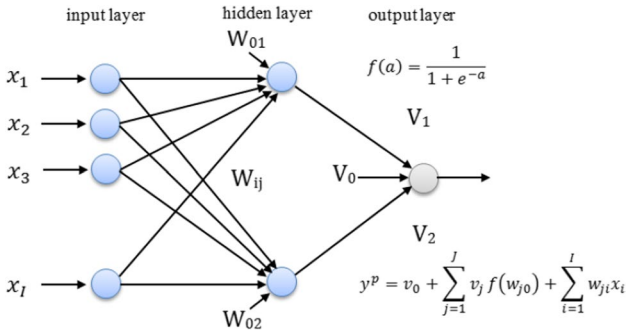

Fig. 3 Schematic multi-layer perceptron (MLP) network

operator, three types of transpositions, and three types of combinations have been utilized.

Artificial neural network model

ANN model can be categorized based on the network structure and learning algorithm (Fig. 3). One of the most applicable fields of ANN is multi-layer perceptron (MLP), based on the back-propagation training algorithm (Haykin 1998). Different nodes are set in groups; namely, layers (commonly, input layer, hidden layer, and output layer) form MLP. The number of input and output variables should be determined, to calculate the number of nodes in input and output layers. Due to simplify calculations, the measured input and output variables can be normalized between 0 and 1. The nodes in consecutive layers represent weighted connections w and v. The performance of each node is the weighted sum of its inputs and filtering them by the means of the transfer function. In this paper, the sigmoidal function has been used in the hidden layer and its equation is as below.

$$f(a) = \frac{1}{1 + e^{-a}} \tag{5}$$

$$y^p = v_0 + \sum_{j-1}^{J} v_j f\left(w_{j0} + \sum_{i-1}^{I} w_{ji} x_i \right) \tag{6}$$

where y^p represents the obtained value of the target, j represents the number of hidden nodes, i represents the number of input nodes, and $w_{j,i,}$ and v_j represent the vectors of weights linking special nodes in the sequential layers.

Structure of WANN model

The developed WANN models are the combination of ANN model and wavelet transform in which the data are pre-processed using wavelet transform and then entered into the ANN model. The main advantage of this approach is increasing the accuracy of the ANN models. WANN models include two stages. In the first stage, wavelet analysis decomposes the main monthly precipitation time series into two main sub-series including estimations (A) and details (D), using the high-pass and low-pass filters, respectively. It should be noted that in wavelet analysis the mother wavelet should be selected properly and according to the main structure of time series (Nourani et al. 2009b; Adamowski and Sun 2010; Danandeh Mehr et al. 2013). Due to the high similarity of precipitation time series with Daubechies mother wavelet, in this study db4 mother has been used for decomposing precipitation time series. Also, for all mentioned stations, two dyadic decomposition levels were determined based on Eq. 7 for considered precipitation time series.

$$L = \mathrm{INT}\left[\log(N)\right] \tag{7}$$

where L represents the number of decomposition levels, INT represents integer operator, N represents the length of the initial time series. All subseries were used as inputs to the ANN models because an averaging or optimizing selection of only certain sub-series would have been a diminutive approach. All sub-series coefficients are equally important and contain information about the original time series (Adamowski and Sun 2010). For example, Fig. 4 indicates two-level detailed sub-signals and second estimation sub-signal computed from two levels DWT decomposition of monthly precipitation time series of Ardabil station.

Structure of WGEP model

The wavelet genetic expression programming (WGEP) is the coupled of DWT and GEP. Like WANN, in the WGEP model, inputs time series will be decomposed by DWT, obtained using chosen mother wavelet function (db4). In the WGEP model, the main inputs time series should be decomposed into multi sub-series, considered as the inputs of the GEP approach. These sub-series represent temporal features of the main time series and were applied as input data to the GEP approach for forecasting the precipitation time series (Fig. 5).

Models development

In this section, the structure of developed precipitation models through ANN, GEP, WANN, WGEP will be considered and described. Mentioned models considered the previous 225 months precipitation time series of four stations including Ardabil, Khalkhal, Meshginshahr, and Parsabad. For the determination of the most appropriate input patterns, autocorrelation function (ACF) and partial autocorrelation function (PACF) were used, for example, these statistical tools presented in Fig. 6 for Ardabil precipitation time series. Also, the mathematical structures of five developed input combinations were considered as follows:

$$\text{(i)}: \; P_{(t)} = P_{(t-1)} \tag{8}$$

$$\text{(ii)}: \; P_{(t)} = P_{(t-1)}, P_{(t-2)}, P_{(t-3)} \tag{9}$$

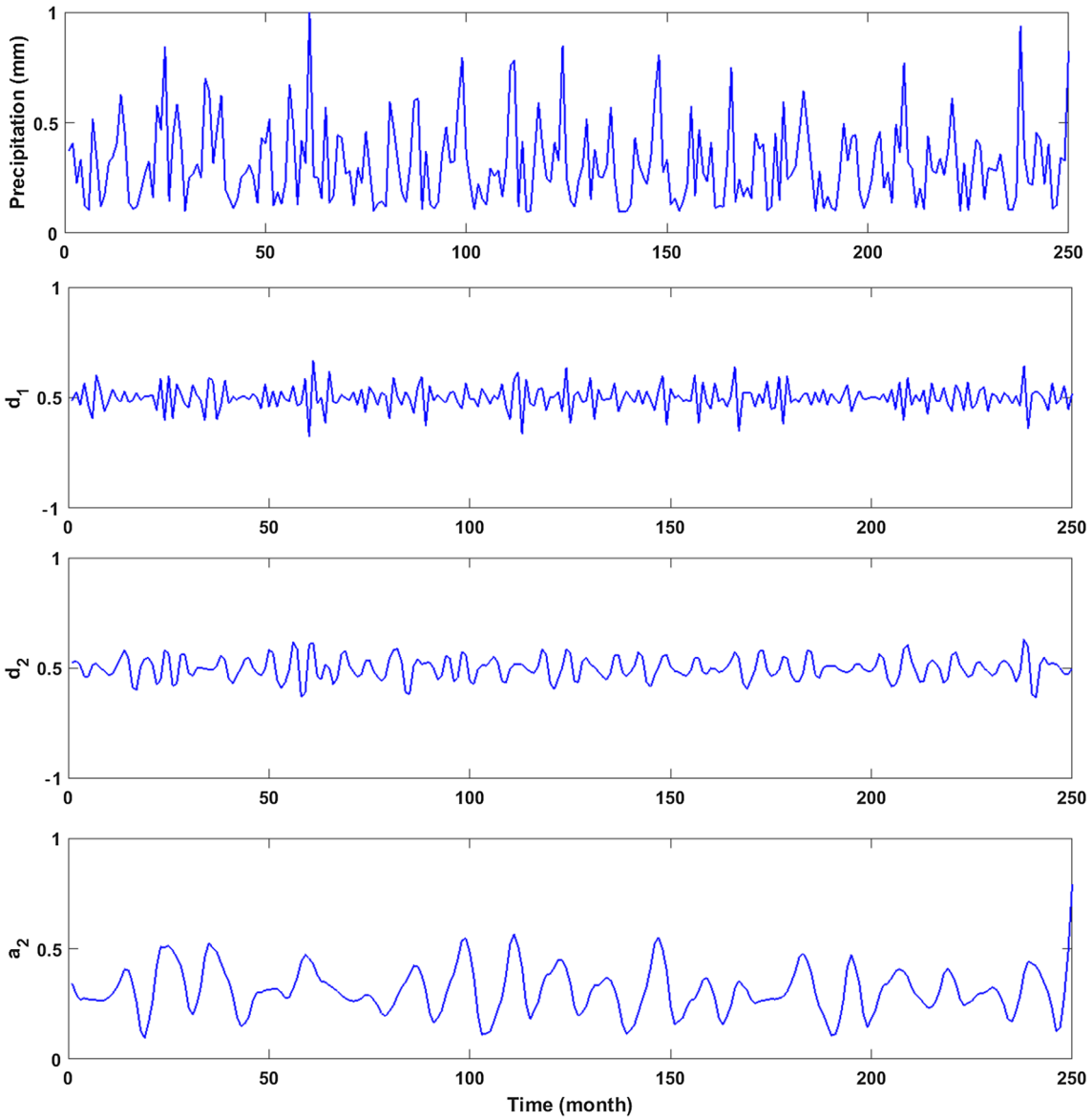

Fig. 4 Decomposition of precipitation time series of Ardabil station

$$\text{(iii)}: \; P_{(t)} = P_{(t-1)}, P_{(t-2)}, P_{(t-3)}, P_{(t-4)}, P_{(t-5)} \tag{10}$$

$$\text{(iv)}: \; P_{(t)} = P_{(t-1)}, P_{(t-2)}, P_{(t-3)}, P_{(t-4)}, P_{(t-5)}, P_{(t-6)}, P_{(t-7)}, P_{(t-8)} \tag{11}$$

$$\begin{aligned}\text{(v)}: P_{(t)} = & P_{(t-1)}, P_{(t-2)}, P_{(t-3)}, P_{(t-4)}, P_{(t-5)}, \\ & P_{(t-6)}, P_{(t-7)}, P_{(t-8)}, P_{(t-9)}, \\ & P_{(t-10)}, P_{(t-11)}, P_{(t-12)}\end{aligned} \tag{12}$$

where P represents mean daily precipitation, $t-1$, $t-2$, …, and $t-12$ represents the current month, 1-month delay, 2 months delay, …, and 12 months delay. It should be mentioned that before application of precipitation data as inputs to these models, data are normalized using the minimum–maximum method in the interval 0.1 through 1 conducted using Eq. 13 (Sajikumar and Thandaveswara 1999).

$$X_n = 0.1 + 0.8 \times \left[\frac{X}{X_{\max} - X_{\min}}\right] \tag{13}$$

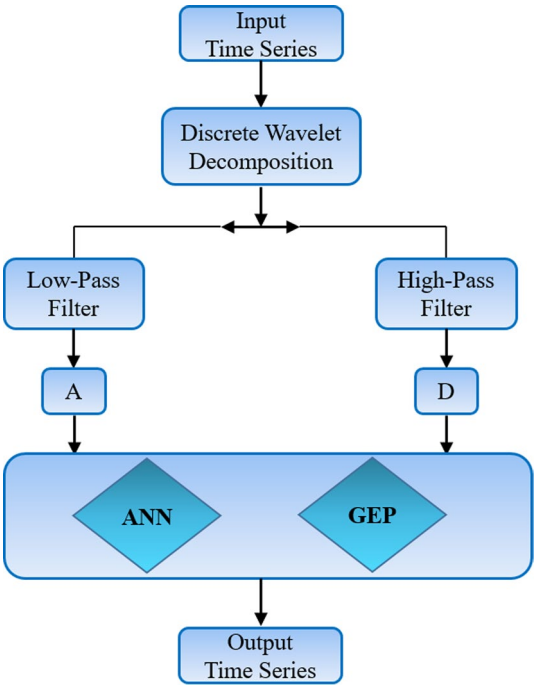

Fig. 5 Schematic view of WANN and WGEP models

where X_n represents the normalized value, $X_{\max}$ represents the maximum value of initial data, and $X_{\min}$ represents the minimum value of initial data.

Evaluation of model performance

Several statistical tests were used for evaluating the performance of applied models. After the calibration of model structures, the performance was evaluated according to the statistical measures of goodness of fit. The coefficient of determination (R^2) and the root mean square error (RMSE) were used for indicating the goodness of fit between the target and output values. The coefficient of determination (R^2) is used for determining the amount of correlation between the target and output values. The higher the R^2 value, the model performs more efficiently. R^2 is determined as follows:

$$R^2 = \frac{\sum_{j-1}^{N} \left(\hat{y}_i - \bar{y}_i\right)^2}{\sum_{j-1}^{N} \left(y_i - \bar{y}_i\right)^2} \tag{14}$$

$$\bar{y}_i = \frac{1}{N} \sum_{i-1}^{N} y_i \tag{15}$$

where $\bar{y}_i$ represents the mean value corresponding to N data, N represents the number of data points applied, y_i represents the target precipitation value, and $\hat{y}_i$ represents the output precipitation value.

Also, the variance of error was evaluated using root mean square error (RMSE) as following:

$$\text{RMSE} = \sqrt{\frac{\text{SEE}}{N}} \tag{16}$$

where SSE is the abbreviation of sum squared error, and it can be calculated as follows:

$$\text{SEE} = \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2 \tag{17}$$

It should be noted that if the value of RMSE is closer to zero, then the model will be more efficient (Nourani et al. 2009a). Moreover, to evaluate the performance of data intelligence models objectively, scatterplot, time series plot, boxplot, and Taylor diagram were utilized.

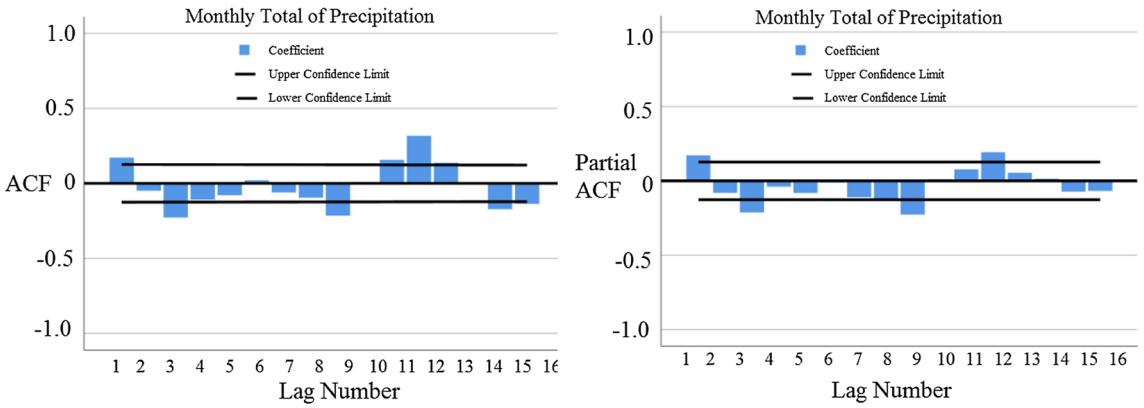

Fig. 6 Autocorrelation function (ACF) and partial autocorrelation function (PACF) of monthly precipitation data in Ardabil station

Table 2 Statistical measures of ANN, GEP, WANN, and WGEP results at validation period for four considered stations

Station	Inputs	ANN			GEP		WANN			WGEP	
		NH†	RMSE	R^2	R^2	RMSE	NH†	RMSE	R^2	R^2	RMSE
AL	1	7	23.06	0.12	0.09	40.34	8	17.90	0.21	0.31	16.80
	2	6	24.41	0.11	0.17	29.32	7	13.62	0.56	0.57	15.36
	3	10	25.72	0.10	0.12	31.55	2	9.34	0.76	0.60	14.13
	4	7	21.24	0.16	0.18	20.33	3	7.55	0.85	0.74	13.06
	5	9	**19.47**	**0.29**	**0.22**	**19.82**	2	**7.13**	**0.88**	**0.79**	**12.48**
KL	1	9	29.17	0.15	0.29	28.11	9	17.19	0.32	0.36	16.82
	2	8	33.02	0.06	0.31	23.05	6	10.67	0.70	0.41	15.32
	3	3	21.35	0.29	**0.36**	**19.00**	1	9.30	0.78	**0.83**	**8.32**
	4	3	27.92	0.15	0.26	26.79	6	9.27	0.80	0.69	10.41
	5	1	**17.00**	**0.34**	0.28	27.80	1	**6.04**	**0.91**	0.63	10.89
MR	1	10	45.06	0.09	0.13	33.20	9	18.14	0.60	0.58	18.24
	2	8	31.65	0.11	0.18	31.17	8	13.98	0.76	0.79	11.06
	3	3	**23.40**	**0.19**	**0.23**	**24.01**	9	13.08	0.69	0.73	11.47
	4	2	23.99	0.14	0.19	28.82	3	**6.57**	**0.92**	**0.81**	**10.18**
	5	6	38.09	0.06	0.11	36.52	2	9.51	0.82	0.68	12.03
PD	1	9	29.84	0.12	0.17	21.07	1	16.42	0.66	0.40	18.30
	2	9	31.02	0.08	0.14	33.58	4	19.07	0.59	0.57	17.74
	3	2	**25.59**	**0.21**	**0.19**	**23.79**	3	12.82	0.73	0.61	15.49
	4	2	28.06	0.18	0.07	35.28	2	**8.56**	**0.87**	**0.82**	**9.94**
	5	9	30.21	0.10	0.07	34.91	2	10.21	0.76	0.77	12.02

NH† number of hidden layers

Bold values refer to the best value of each criteria

Results and discussion

The main motivation of the current study was the exploration of a reliable data intelligence model for precipitation forecasting. This section covers the evaluation and assessment of the developed forecasting models. The statistical measures of ANN, GEP, WANN, and WGEP models for validation, corresponding to 5 input combinations (i.e., 8–12 Eqs.) that incorporated the correlated lead times of the historical memory precipitation, are reported in Table 2. The performance of designed models shows that in Ardabil, Khalkhal, Meshginshahr, and Parsabad stations, the performance of ANN and GEP models is very poor for various five input combinations, but pre-processing applied by wavelet transform has improved the performance of all models; so that in Ardabil station corresponding to input (v) (the best input combination in this station), the values of ANN and GEP models are equal to 0.29 and 0.22 in terms of R^2, respectively, while the values of WANN and WGEP models at this station corresponding to input (v) are equal to 0.88 and 0.79 for the R^2 index, respectively. It is because WANN and WGEP models record long, medium, and low patterns in themselves, while ANN and GEP models are only able to record low time-series patterns. The superiority of applicability and performance of WANN and WGEP models to ANN and GEP models in three other stations for various input combinations are presented in Table 2 because of the pre-processing applied by wavelet analysis on the input variables of the reference model resulting in the better learning of network.

Among different input combinations, input (v) has indicated the best performance corresponding to four proposed models in Ardabil station that the best performance is also related to the WANN model for which the values of R^2 and RMSE are equal to 0.88 and 7.13 mm, respectively. Therefore, it is concluded that the more the value of delay in this station is, the applicability of models will be higher. This is also evidenced by the necessity of long historical information to build the learning process of the forecasting models at this station.

Although in Khakhal station, increasing produced delay, the results of ANN and WANN models have been improved (input (v)), this delay has not affected the performance and applicability of GEP and WGEP models. According to the results, input combination (iii) has the best performance in terms of R^2 and RMSE indices for these two models. In the Khakhal station, the WANN model has the best efficiency with $R^2 = 0.91$ and RMSE = 6.04 mm.

The conditions of Meshginshahr and Parsabad stations were observed to be similar to each other in terms of the performance of reference models so that for ANN and GEP

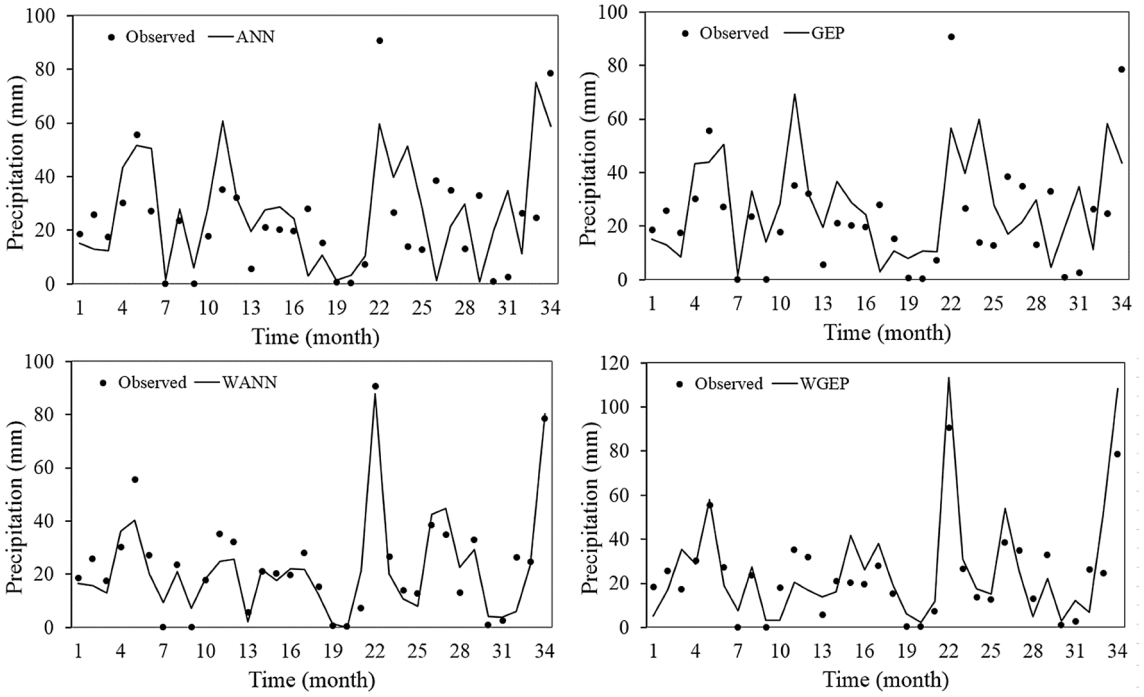

Fig. 7 Predicted and observed precipitation time series for various applied models at Ardabil station

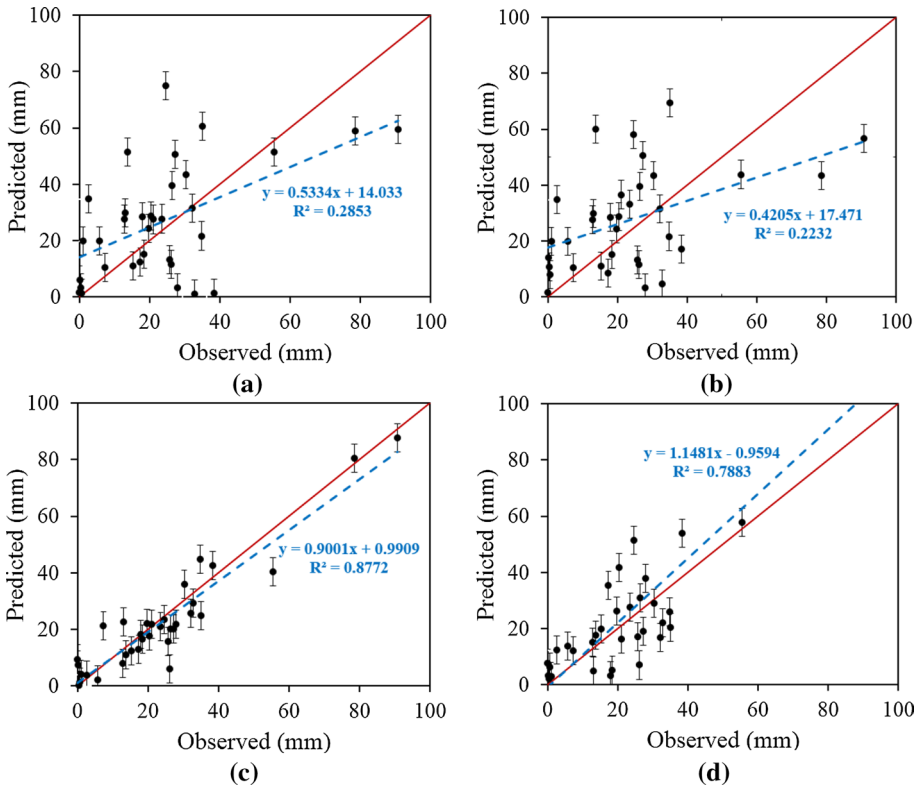

Fig. 8 Scatterplots of validation phase at Ardabil station for **a** ANN, **b** GEP, **c** WANN, and **d** WGEP models

models, input (iii) indicates the best performance and for WANN and WGEP models; input (iv) has the best performance. Among intelligent models, the best performance in Meshginshahr station is related to the WANN model for which the values of R^2 and RMSE indices are equal to 0.92 and 6.57 mm, respectively; also, this model indicates the

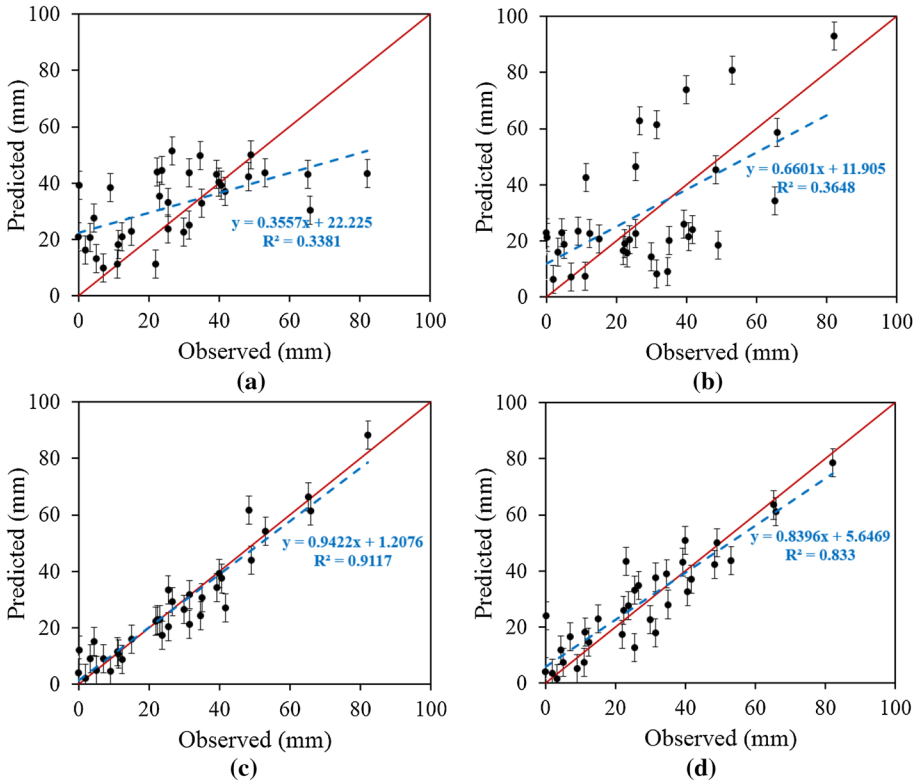

Fig. 9 Scatterplots of validation phase at Khalkhal station for **a** ANN, **b** GEP, **c** WANN, and **d** WGEP models

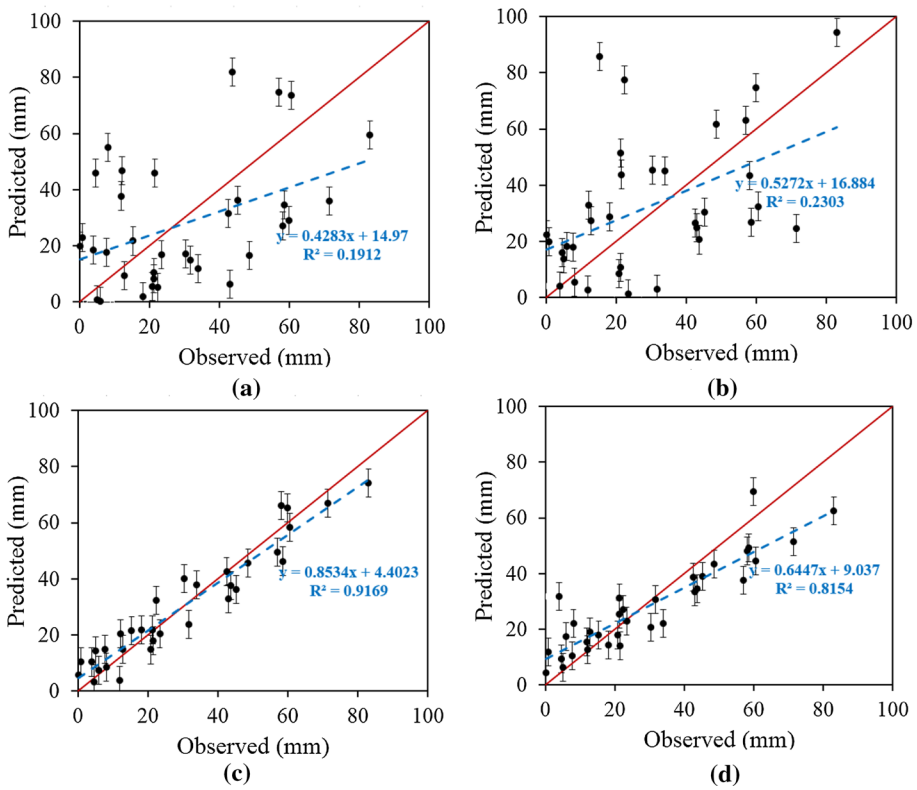

Fig. 10 Scatterplots of validation phase at Meshginshahr station for **a** ANN, **b** GEP, **c** WANN, and **d** WGEP models

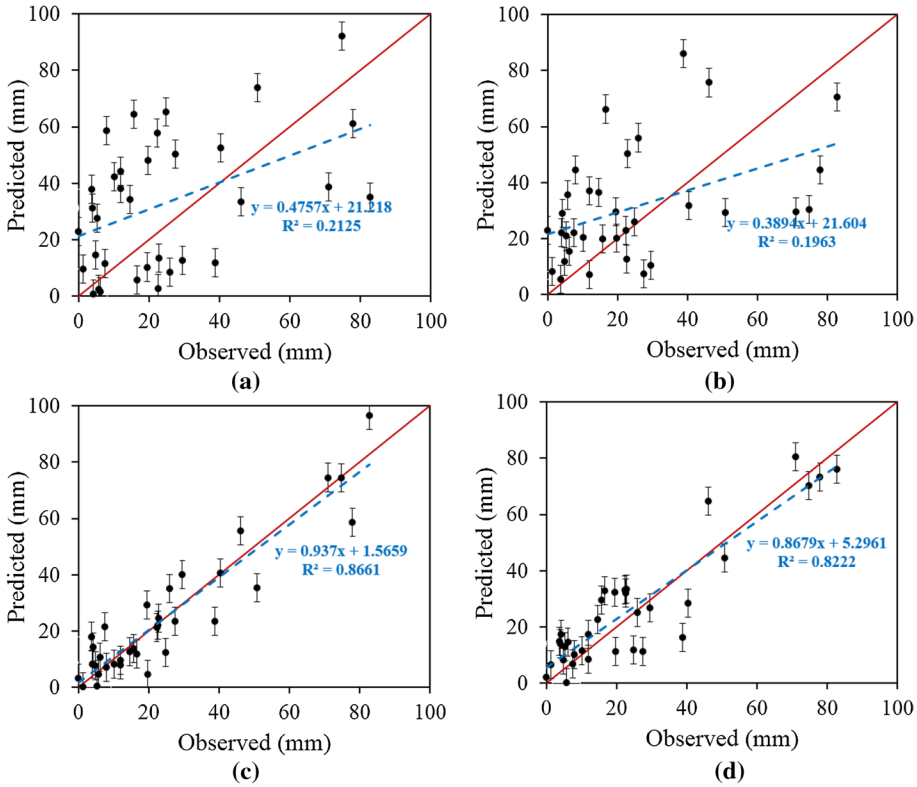

Fig. 11 Scatterplots of validation phase at Parsabad station for **a** ANN, **b** GEP, **c** WANN, and **d** WGEP models

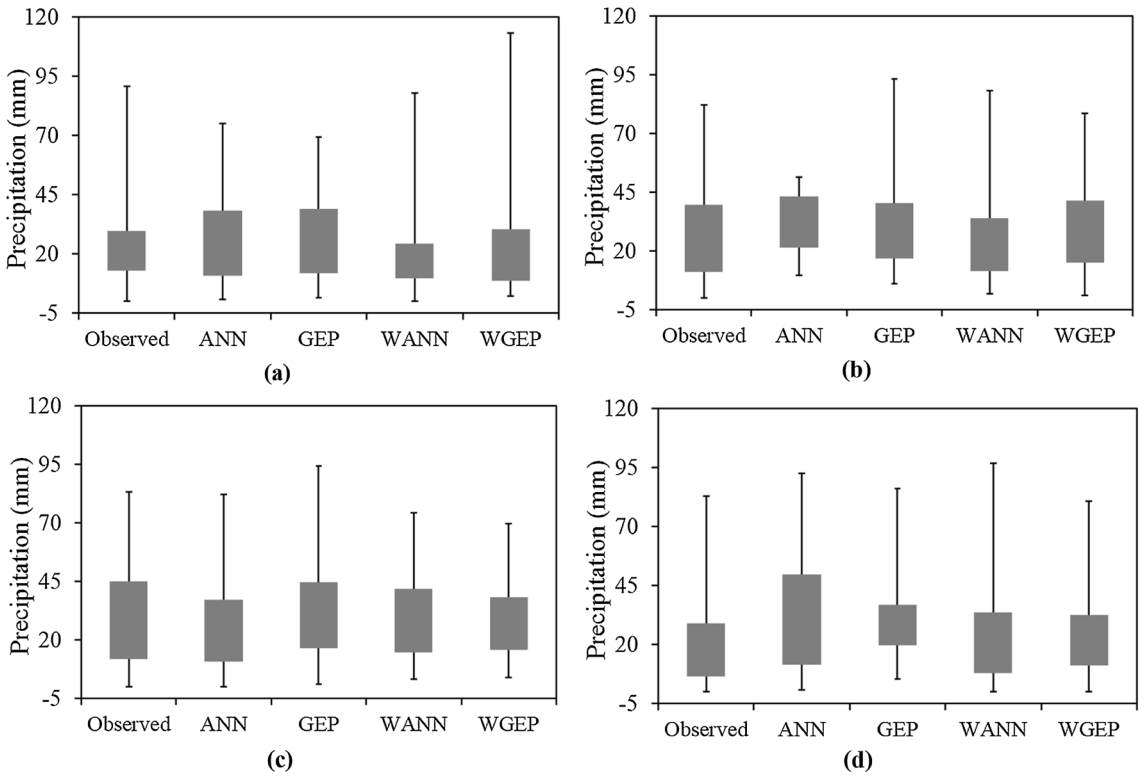

Fig. 12 Boxplots of the observed rainfall compared with simulated rainfall using ANN, GEP, WANN, and WGEP models for different stations **a** Ardabil, **b** Khalkhal, **c** Meshginshahr, and **d** Parsabad, during the test period

best efficiency among intelligent models in Parsabad station that the values of R^2 and RMSE are equal to 0.87 and 8.56 mm, respectively.

The observed and forecasted precipitation time series for different models at Ardabil station is shown in Fig. 7. Furthermore, Fig. 8 shows the scatterplots of different models in the Ardabil station during the validation phase. It should be mentioned that the solid and dotted lines shown in Fig. 8 are the line 1:1 and trend line of the data, respectively. In Ardabil station, according to Fig. 7, the WANN model has forecasted most peak precipitations lower than real values; in other words, peak precipitations have been underestimated. On the contrary, as it is indicated in Fig. 7, the ANN model has forecasted most precipitation values higher than observed values, and generally, they have been overestimated. Unlike two other mentioned models, GEP and WGEP models have not overall forecasted peak values of precipitations well (Figs. 7, 8).

At Khalkhal station, according to the scatters drawn in Fig. 9 it can be concluded that the hybrid wavelet–artificial neural network model has performed better than other models, so that it has been able to forecast extreme values, but in this model with increasing precipitation values, overestimation has occurred. Also, according to the other scatters related to Khalkhal station, it can be noted that the performance of ANN and GEP models is not appropriate; also in WGEP until the values of precipitation are low, forecasted values are appropriate. In Meshkinshahr station, the scatters

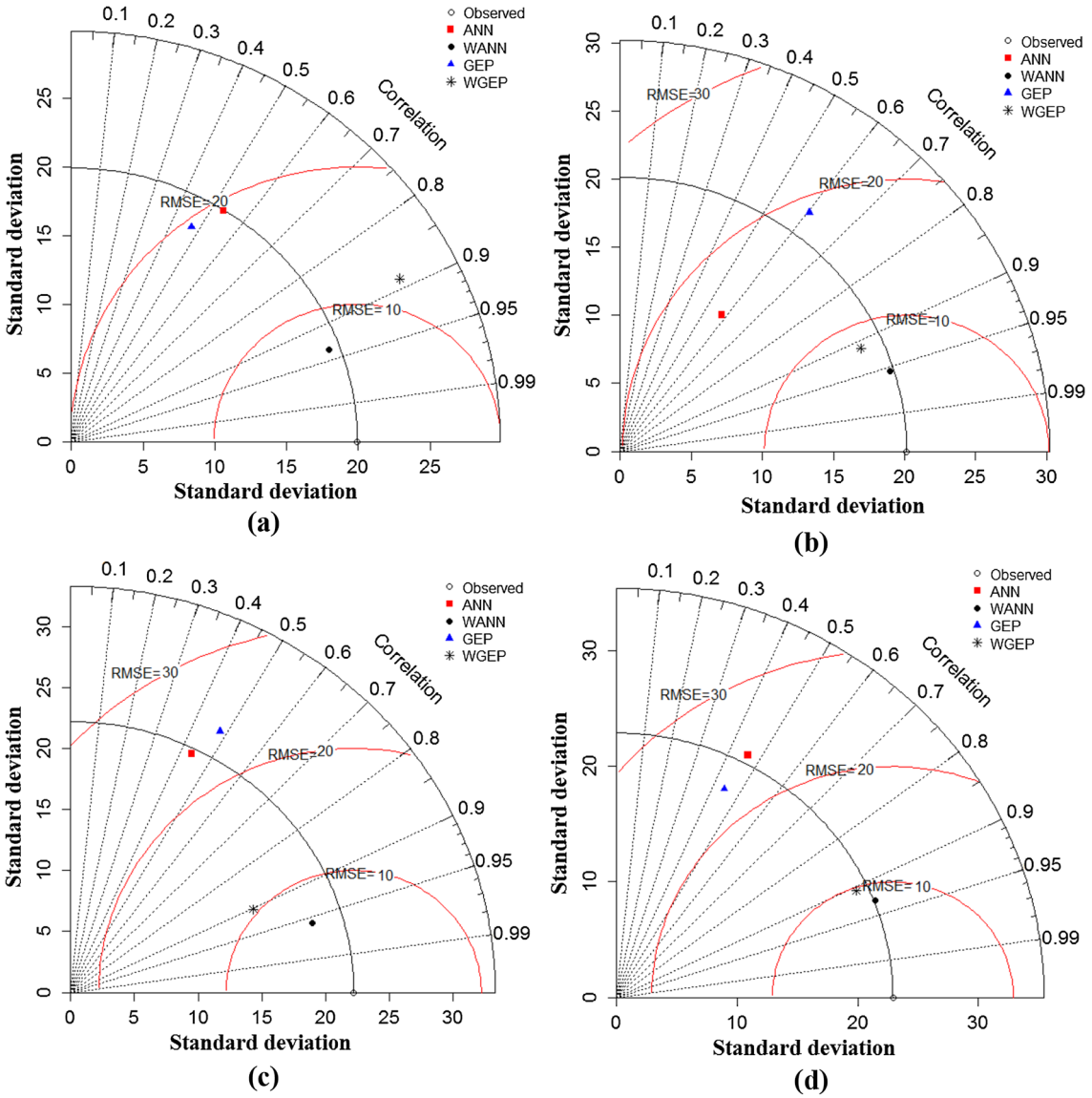

Fig. 13 Graphical presentation of the performance (Taylor diagram) for different models, namely ANN, GEP, WANN, and WGEP for different stations **a** Ardabil, **b** Khalkhal, **c** Meshginshahr, and **d** Parsabad, during the test period

related to GEP and ANN models indicate that the performance of these two models in the estimation of precipitation is not appropriate, while the scatters related to WGEP and WANN models show that in the maximum values of precipitation, overestimation has been occurred (Fig. 10).

At Parsabad station, the best model was obtained using WANN model as it could forecast the extreme values, accurately. However, this model forecasting was not performed well, and over-estimation has occurred. It should be noted that in the WGEP model, for the low values of precipitation the accuracy of forecasting is nearly appropriate, but with increasing the values of precipitation, forecasted values are not appropriate (Fig. 11).

Based on the chaotic behavior of hydrological variables (i.e., precipitation), the cross-validation of models considering the distribution of observed and simulated methods is important. In this regard, boxplots are useful to demonstrate the forecasts of applied models compared to the original time series. Figure 12 shows the boxplots to indicate the degree of overall spread of forecasted and observed precipitation data with the values of the respective quartiles. Figure 12, at Ardabil station, shows the WANN forecasted precipitation median, quintiles, and data ranges somewhat similar to the observed values. In this station, WGEP has been found to over-estimate the higher precipitation values. In the Khalkhal and Parsabad stations, WGEP model forecasts closely resembled the observed precipitation with a slightly lower interquartile range. Regarding the Meshginshahr station, it was found that ANN model showed lower and median values closer to observed data and the forecasting of other models have some fluctuations to mimic the observed values.

The performance of various applied data intelligence models was also evaluated using graphical presentation (i.e., Taylor diagram) for checking the agreement between models and observations, shown in Fig. 13. Some model statistics including correlation centered RMSE difference and standard deviation are presented using the Taylor diagram (Taylor 2001). The visualization of Fig. 13 showed that WANN model results were closer to the reference point on the standard deviation axis than the ANN, GEP, and WGEP model results that reaffirm the better performance and accuracy of the WANN model.

Conclusion

In this study, the comparison of four data intelligence models (i.e., ANN, GEP, WANN, and WGEP) was investigated for monthly precipitation forecasting. Historical datasets were obtained from four stations (i.e., Ardabil, Khalkhal, Meshginshahr, and Parsabad) in Iran and were used for this investigation. The statistical correlation analysis was adopted for the determination of the lag numbers. The models were evaluated using statistical and graphical indicators. The forecasting models were initiated using five different input combinations including different lags information. The results confirmed that among various input combinations, at Ardabil station, input (v) performed better than four other inputs in which the WANN model was the most applicable one. Furthermore, at the other three stations, the WANN model was more efficient than the standalone models. Similarly, in the researches performed by Nourani et al. (2009a), Arab Amiri et al. (2016), Zhang et al. (2018) and Ghamariadyan and Imteaz (2021), the WANN model was identified as the superior model compared to the other models for monthly rainfall prediction. The forecasting performance was obtained for Khalkhal, Meshginshahr, and Parsabad stations corresponding to (v), (iv), and (iv) input combinations, respectively. The results showed that the accuracy of the ANN and GEP model is not reliably with confident level of accuracy. It can be concluded that the combination of the wavelet method with ANN and GEP models results in improvements in the forecasting of precipitation time series. This is due to the decomposing the main time series into multi-scale sub-monthly signals. For example, at the Khalkhal station, the value of R^2 obtained from the ANN model in input (v) was equal to 0.34 while this index was obtained 0.91 using the WANN model. Also, the values of R^2 in the Meshginshahr station determined by GEP and WGEP for input (ii) were 0.07 and 0.81, respectively. Another result of this study is the forecasting of peak precipitation values. It can be claimed that ANN model overestimated the forecasting of precipitation time series, while these time series were under-estimated by WANN model, whereas GEP and WGEP were not overall appropriate for forecasting the peak time series. These methodologies can be simply applied for enabling the design of similar problems in the field of water resources engineering and applied as a hybrid alternative to improve other hydrological models such as regionalization watersheds techniques.

Acknowledgments The authors wish to thank all who assisted in conducting this work.

Authors' contributions MRN took part in writing the paper and original draft preparation. SA involved in conceptualization and methodology. HS participated in visualization and investigation. JR involved in modeling. ZMY took part in reviewing and editing.

Funding The authors did not receive support from any organization for the submitted work.

Availability of data and materials Not applicable.

Declarations

Conflict of interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate The authors declare that they voluntarily agree to participate in this research.

Consent to publish Not applicable.

References

Abbot J, Marohasy J (2012) Application of artificial neural networks to rainfall forecasting in Queensland, Australia. Adv Atmos Sci 29:717–730. https://doi.org/10.1007/s00376-012-1259-9

Abdollahi S, Raeisi J, Khalilianpour M et al (2017) Daily mean streamflow prediction in perennial and non-perennial rivers using four data driven techniques. Water Resour Manag 31:4855–4874. https://doi.org/10.1007/s11269-017-1782-7

Adamowski JF (2008a) River flow forecasting using wavelet and cross-wavelet transform models. Hydrol Process 22:4877–4891. https://doi.org/10.1002/hyp.7107

Adamowski JF (2008b) Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis. J Hydrol 353:247–266. https://doi.org/10.1016/j.jhydrol.2008.02.013

Adamowski J, Sun K (2010) Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. J Hydrol 390:85–91. https://doi.org/10.1016/j.jhydrol.2010.06.033

Ahani A, Mousavi Nadoushani SS, Moridi A (2018) A feature weighting and selection method for improving the homogeneity of regions in regionalization of watersheds. Hydrol Process 32:2084–2095

Ahani A, Mousavi Nadoushani SS, Moridi A (2020a) Regionalization of watersheds based on the concept of rough set. Nat Hazards 104:883–899

Ahani A, Mousavi Nadoushani SS, Moridi A (2020b) Simultaneous regionalization of gauged and ungauged watersheds using a missing data clustering method. J Hydrol Eng 25:04020015

Ahani A, Nadoushani SSM, Moridi A (2020c) Regionalization of watersheds by finite mixture models. J Hydrol 583:124620

Ali M, Parsad R, Xiang Y, Mundher Yaseen Z (2020) Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. J Hydrol 584:124647. https://doi.org/10.1016/j.jhydrol.2020.124647

Arab Amiri M, Amerian Y, Mesgari MS (2016) Spatial and temporal monthly precipitation forecasting using wavelet transform and neural networks, Qara-Qum catchment. Iran Arab J Geosci 9:421. https://doi.org/10.1007/s12517-016-2446-2

Cannas B, Fanni A, See L, Sias G (2006) Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. Phys Chem Earth Parts ABC 31:1164–1171. https://doi.org/10.1016/j.pce.2006.03.020

Chou C (2011) A threshold based wavelet denoising method for hydrological data modelling. Water Resour Manag 25:1809–1830. https://doi.org/10.1007/s11269-011-9776-3

Christodoulou SE, Kourti E, Agathokleous A (2017) Waterloss detection in water distribution networks using wavelet change-point detection. Water Resour Manag 31:979–994. https://doi.org/10.1007/s11269-016-1558-5

Dabhi VK, Chaudhary S (2014) Hybrid wavelet-postfix-GP model for rainfall prediction of Anand Region of India. Adv Artif Intell 3(3–3):3. https://doi.org/10.1155/2014/717803

Danandeh Mehr A, Kahya E, Olyaie E (2013) Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. J Hydrol 505:240–249. https://doi.org/10.1016/j.jhydrol.2013.10.003

Danandeh Mehr A, Nourani V (2018) Season algorithm-multigene genetic programming: a new approach for rainfall-runoff modelling. Water Resour Manag 32:2665–2679. https://doi.org/10.1007/s11269-018-1951-3

Dawkins LC, Osborne JM, Economou T, Darch GJC, Stoner OR (2022) The advanced meteorology explorer: a novel stochastic, gridded daily rainfall generator. J Hydrol 607:127478. https://doi.org/10.1016/j.jhydrol.2022.127478

DelSole T, Shukla J (2012) Climate models produce skillful predictions of Indian summer monsoon rainfall. Geophys Res Lett. https://doi.org/10.1029/2012GL051279

Deo RC, Şahin M (2015) Application of the Artificial Neural Network model for prediction of monthly Standardized Precipitation and Evapotranspiration Index using hydrometeorological parameters and climate indices in eastern Australia. Atmos Res 161–162:65–81. https://doi.org/10.1016/j.atmosres.2015.03.018

Dökmen F, Aslan Z (2013) Evaluation of the parameters of water quality with wavelet techniques. Water Resour Manag 27:4977–4988. https://doi.org/10.1007/s11269-013-0454-5

Everingham Y, Baillie C, Inman-Bamber G, Baillie J (2008) Forecasting water allocations for Bundaberg sugarcane farmers. Clim Res 36:231–239. https://doi.org/10.3354/cr00743

Falayi EO, Adepitan JO, Adewole AT, Roy-Layinde TO (2022) Analysis of rainfall data of some West African countries using wavelet transform and nonlinear time series techniques. J Spat Sci. https://doi.org/10.1080/14498596.2021.2008539

Feng Q, Wen X, Li J (2015) Wavelet analysis-support vector machine coupled models for monthly rainfall forecasting in Arid Regions. Water Resour Manag 29:1049–1065. https://doi.org/10.1007/s11269-014-0860-3

Ferreira C (2006) Gene expression programming: mathematical modeling by an artificial intelligence. Springer, Berlin

Ghamariadyan M, Imteaz MA (2021) A wavelet artificial neural network method for medium-term rainfall prediction in Queensland (Australia) and the comparisons with conventional methods. Int J Climatol 41:1396–1416. https://doi.org/10.1002/joc.6775

Haykin S (1998) Neural networks: a comprehensive foundation, Subsequent edition. Prentice-Hall, Upper Saddle River, 842 p

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. https://doi.org/10.1016/j.jhydrol.2015.04.047

Kisi O, Cimen M (2012) Precipitation forecasting by using wavelet-support vector machine conjunction model. Eng Appl Artif Intell 25:783–792. https://doi.org/10.1016/j.engappai.2011.11.003

Kisi O, Shiri J (2011) Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. Water Resour Manag 25:3135–3152. https://doi.org/10.1007/s11269-011-9849-3

Komasi M, Sharghi S (2016) Hybrid wavelet-support vector machine approach for modelling rainfall–runoff process. Water Sci Technol 73:1937–1953

Kuo C-C, Gan TY, Yu P-S (2010) Wavelet analysis on the variability, teleconnectivity, and predictability of the seasonal rainfall of

Taiwan. Mon Weather Rev 138:162–175. https://doi.org/10.1175/2009MWR2718.1
Lafrenière M, Sharp M (2003) Wavelet analysis of inter-annual variability in the runoff regimes of glacial and nival stream catchments, Bow Lake, Alberta. Hydrol Process 17:1093–1118. https://doi.org/10.1002/hyp.1187
Lane SN (2007) Assessment of rainfall-runoff models based upon wavelet analysis. Hydrol Process 21:586–607. https://doi.org/10.1002/hyp.6249
Lim E-P, Hendon HH, Hudson D et al (2009) Dynamical forecast of Inter–El Niño variations of tropical SST and Australian spring rainfall. Mon Weather Rev 137:3796–3810. https://doi.org/10.1175/2009MWR2904.1
Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11:674–693. https://doi.org/10.1109/34.192463
Mehdizadeh S, Behmanesh J, Khalili K (2018) New approaches for estimation of monthly rainfall based on GEP-ARCH and ANN-ARCH hybrid models. Water Resour Manag 32:527–545. https://doi.org/10.1007/s11269-017-1825-0
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. https://doi.org/10.1016/j.jhydrol.2013.08.035
Miao J, Liu G, Cao B et al (2014) Identification of strong karst groundwater runoff belt by cross wavelet transform. Water Resour Manag 28:2903–2916. https://doi.org/10.1007/s11269-014-0645-8
Mwale D, Gan TY (2005) Wavelet analysis of variability, teleconnectivity, and predictability of the September–November East African rainfall. J Appl Meteorol 44:256–269. https://doi.org/10.1175/JAM2195.1
Mwale D, Gan TY, Shen SSP, Shu TT, Kim KM (2007) Wavelet empirical orthogonal functions of space-time-frequency regimes and predictability of Southern Africa summer rainfall. J Hydrol Eng 12:513–523. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(513)
Nourani V, Alami MT, Aminfar MH (2009a) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Eng Appl Artif Intell 22:466–472. https://doi.org/10.1016/j.engappai.2008.09.003
Nourani V, Komasi M, Mano A (2009b) A multivariate ANN-wavelet approach for rainfall-runoff modeling. Water Resour Manag 23:2877. https://doi.org/10.1007/s11269-009-9414-5
Partal T, Cigizoglu HK (2009) Prediction of daily precipitation using wavelet—neural networks. Hydrol Sci J 54:234–246. https://doi.org/10.1623/hysj.54.2.234
Partal T, Kişi Ö (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. J Hydrol 342:199–212. https://doi.org/10.1016/j.jhydrol.2007.05.026
Rathinasamy M, Agarwal A, Parmar V et al (2017) Partial wavelet coherence analysis for understanding the standalone relationship between Indian Precipitation and Teleconnection patterns
Rezaei M, Mousavi SF, Moridi A et al (2021) A new hybrid framework based on integration of optimization algorithms and numerical method for estimating monthly groundwater level. Arab J Geosci 14:1–15
Roushangar K, Alizadeh F, Nourani V (2018) Improving capability of conceptual modeling of watershed rainfall–runoff using hybrid wavelet-extreme learning machine approach. J Hydroinform 20:69–87. https://doi.org/10.2166/hydro.2017.011
Sajikumar N, Thandaveswara BS (1999) A non-linear rainfall–runoff model using an artificial neural network. J Hydrol 216:32–55. https://doi.org/10.1016/S0022-1694(98)00273-X
Salih SQ, Sharafati A, Ebtehaj I, Sanikhani H, Siddique R, Deo RC, Bonakdari H, Shahid Sh, Mundher Yaseen Z (2020) Integrative stochastic model standardization with genetic algorithm for rainfall pattern forecasting in tropical and semi-arid environments. Hydrol Sci J 65:1145–1157. https://doi.org/10.1080/02626667.2020.1734813
Schaefli B, Maraun D, Holschneider M (2007) What drives high flow events in the Swiss Alps? Recent developments in wavelet spectral analysis and their application to hydrology. Adv Water Resour 30:2511–2525. https://doi.org/10.1016/j.advwatres.2007.06.004
Schepen A, Wang QJ, Robertson DE (2012) Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. J Geophys Res Atmos. https://doi.org/10.1029/2012JD018011
Sezen C, Partal T (2017) A wavelet transformation-genetic algorithm-artificial neural network combined model for precipitation forecasting. Eurasia Proc Sci Technol Eng Math 1:372–378
Shafaei M, Adamowski J, Fakheri-Fard A et al (2016) A wavelet-SARIMA-ANN hybrid model for precipitation forecasting. J Water Land Dev 28:27–36. https://doi.org/10.1515/jwld-2016-0003
Shoaib M, Shamseldin AY, Melville BW, Khan MM (2015) Runoff forecasting using hybrid Wavelet Gene Expression Programming (WGEP) approach. J Hydrol 527:326–344. https://doi.org/10.1016/j.jhydrol.2015.04.072
Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmosph 106(D7):7183–7192
Venkata Ramana R, Krishna B, Kumar SR, Pandey NG (2013) Monthly rainfall prediction using wavelet neural network analysis. Water Resour Manag 27:3697–3711. https://doi.org/10.1007/s11269-013-0374-4
Wang Y, Liu J, Li R, Suo X, Lu E (2022) Medium and long-term precipitation prediction using wavelet decomposition-prediction-reconstruction model. Water Resour Manag. https://doi.org/10.1007/s11269-022-03063-x
Wu J, Liu M, Jin L (2010) A hybrid support vector regression approach for rainfall forecasting using particle swarm optimization and projection pursuit technology. Int J Comput Intell Appl 9:87–104
Wu X, Zhou J, Yu H, Liu D, Xie K, Chen Y, Hu J, Sun H, Xing F (2021) The development of a hybrid wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis. Atmosphere. https://doi.org/10.3390/atmos12010074
Zeynoddin M, Bonakdari H, Azari A et al (2018) Novel hybrid linear stochastic with non-linear extreme learning machine methods for forecasting monthly rainfall a tropical climate. J Environ Manag 222:190–206. https://doi.org/10.1016/j.jenvman.2018.05.072
Zhang S, Chang T, Lin D (2018) A preliminary study on a hybrid wavelet neural network model for forecasting monthly rainfall. Eurasia J Math Sci Technol Educ 14:1747–1757. https://doi.org/10.29333/ejmste/85119