Theoretical and Applied Climatology (2023) 151:1–11
https://doi.org/10.1007/s00704-022-04143-z

ORIGINAL PAPER

Regionalization of rainfall intensity–duration–frequency (IDF) curves with L-moments method using neural gas networks

Mohammad Reza Mahmoudi[1] · Saeid Eslamian[1] · Saeid Soltani[2] · Moein Tahanian[1]

Received: 8 July 2021 / Accepted: 8 July 2022 / Published online: 5 August 2022

Abstract

Floods are one of the most frequent and destructive natural events which lead to lots of human and financial losses with damage to the houses, farms, roads, and other buildings. Intensity–duration–frequency (IDF) curves are the main and practical tools that have been used for flood control studies, including the design of the water structures. In many cases, there is no measuring device at the desired place, or their information is not helpful if there is any available. In this case, it is not possible to extract these curves through conventional methods. Regionalizing the IDF curves is a method that has solved the issues mentioned in the common methods. In this research, the regionalized IDF curves are extracted in Khuzestan province, Iran using 21 rain gauge stations through L-moments and neural gas networks. Clustering is one of the most effective steps and a prerequisite for regional frequency analysis (RFA) that divides the region and existing stations into hydrologically homogenous regions. In this study, clustering is done using two new models named neural gas (NG) and growing neural gas (GNG) network. Comparing the regional IDF curves with at-site curves, it was found that neural gas network models had a more accurate performance and higher efficiency, so they had the lowest estimate error amount among other models. Also, due to the acceptable difference between regional and at-site curves, the efficiency of L-moments in RFA was evaluated as appropriate.

Keywords Regionalization · IDF curves · Clustering · Neural gas network · L-moments · Khuzestan

1 Introduction

Intensity–duration–frequency (IDF) curve is one of the most common tools in water resources engineering, which can be used as an input in planning and designing, and exploitation of water resources projects. One of the common problems in many countries is the scattered or very weak networks of the required meteorological stations such that their data are considered the main bases for IDF construction. To this problem, a regional analysis of rainfall depth and building the IDF curves have been proposed.

The IDF concept refers to Bernard's efforts in 1932 (Bernard 1932), and a lot of the studies focused on improving the statistical inference methods used in IDF (Bell 1969). One of the noticeable researches in this field is Hasking and Wallis' study (1997) on developing a method for L-moments estimation, probability-weighted moments (PWM) (Greenwood et al. 1979), parametric formulation of IDF relations (Koutsoyiannis et al. 1998), and employing the regional methods like the Index-Flood method. Today, Atlas of IDFs has been built in developed countries. One of the works is the National Oceanic and Atmospheric Administration (NOAA) atlas 14, which was created by American National Weather Services (Perica et al. 2013).

The regional analysis uses the group statistics and characteristics from co-behavioral stations instead of using data only from one station. Several studies using regional methods on the extreme rainfalls suggest that these techniques increasingly reduce the doubts about the estimates resulting

✉ Mohammad Reza Mahmoudi
m.mahmoudi@alumni.iut.ac.ir

Saeid Eslamian
saied@iut.ac.ir

Saeid Soltani
ssoltani@cc.iut.ac.ir

Moein Tahanian
moein.tahanian@ag.iut.ac.ir

[1] Department of Water Engineering, College of Agriculture, Isfahan University of Technology, Isfahan 8415683111, Iran

[2] Department of Natural Resources Engineering, Isfahan University of Technology, Isfahan, Iran

from the at-site view of point (Lee and Maeng 2003). One of the main problems to expand frequency analysis results from one or more stations to one region is the hydrological lack of homogeneity in the region. Despite the suitability of cluster analysis for grouping the hydrological features, the homogeneity of the regions is not completely achieved. So, it is recommended to examine and test the cluster analysis results with the other conventional methods (Rousseeuw 1987).

Soltani et al. (2017), using the characteristics of the rainfall time scale and three variables of average daily rainfall intensity, the standard deviation of daily rainfall intensity, and scale index, drew the regional IDF curves for Khuzestan province and the absolute error of estimates for this method, which was mainly below 25%, and confirmed the results were acceptable.

Using topographic and rainfall characteristics, Alemaw and Chaoka (2016) divided Botswana into three hydrologically homogeneous regions using the K-means clustering model. They also used both gamma and lognormal probability distribution to estimate the depth of rainfall at the return periods up to 100 years for the mentioned areas.

Amin and Shaaban (2004) used generalized extreme values distribution(GEV) and extreme values distribution (EV1) with the least square method in the estimation of distribution parameters for the IDF curves in Peninsular Malaysia.

The IDF regionalization is very beneficial in terms of shortening the steps and required time to perform the calculations as well as providing it for the area and not only for the station.

Identifying the homogeneous regions is usually the most crucial and difficult step and a prerequisite for the frequency analysis hypotheses between the hydrologic frequency analysis stages of the region. This study presents a method based on the neural gas and growing neural gas networks to cluster the hydrological data and determine the homogeneous regions. The neural gas network is one of the types of competitive neural networks and uses an unsupervised teaching method. The network was first introduced by Martinez and Schulten (1991). One of the features of this algorithm is learning the topology or distribution shape of the data space. One of the issues with this algorithm is that it starts working with several elements, which makes the algorithm too slow at first. This problem was solved 4 years later when the growing neural gas network was presented by Fritzke (1994). The number of neurons in the growing neural gas network increases during the learning process regardless of prior knowledge and the governing structure of the inputs.

Abdi et al. (2017) investigated the ability of the growing neural gas network to regionalize the drought index for 40 synoptic stations in Iran, and the results of the heterogeneity evaluation based on L-moments showed the success of this algorithm compared to the other methods in determining the homogeneous sub-regions.

The application of neural gas networks in clustering has also been considered in other scientific disciplines such as robotics (Carlevarino et al. 2000; Ferrer 2014), medicine (Cselényi 2005; Oliveira Martins et al. 2009; Angelopoulou et al. 2015), and economics (Decker 2005; Lisboa et al. 2000). Therefore, this algorithm can be used for clustering and image segmentation.

No studies have been conducted to regionalize the IDF curves using clustering based on neural gas networks. Also, in the field of hydrology and water resources, neural gas networks have not been used so far except for a few cases (Abdi et al. 2017).

After using neural gas networks and other models for clustering, it is required to investigate the formed regions and stations in each area in terms of homogeneity and discordancy. For this purpose, Husking and Wallis tests, which are based on L-moments, are known as the best method for regional analysis.

L-moments were presented by Hosking (1990), and it has been found the great importance and application in many regional analyses. The most crucial applications of L-moments include detecting the homogeneous regions, determining the discordant stations, selecting the appropriate distribution function, and estimating the parameters of distribution functions (Hosking and Wallis 1997). The main advantage of L-moments over ordinary moments is that they can describe a larger range of distributions, and the estimates have less bias. They also work better at displaying outlier events (Rao and Hamed 1997).

In this study, both models of neural gas networks have been used to regionalize the IDF curves in Khuzestan province. The results were then compared using a test based on L-moments combining the conventional clustering models like Ward, K-means, SOM,[1] and FCM.[2] The IDF regional curves were also extracted with the L-moment concepts.

Several studies with the L-moments method have analyzed the frequency of extreme rainfalls and extracted the appropriate distribution function for each region. For instance, Yang et al. (2010) divided the Pearl river basin in China into six homogeneous clusters, and they performed the regional frequency analysis for maximum annual 1, 3, 5, and 7 days of rainfall with L-moments. The goodness of fit tests showed that the distribution functions of PE3,[3] GLO,[4] and GEV fit well in the most areas in the homogeneous regions.

[1] - Self-organizing map.

[2] - Fuzzy C-means.

[3] - Pearson type 3.

[4] - Generalized logistic.

Table 1 Details of meteorological stations

Station number	Station name	Elevation (m)	Latitude	Longitude	MAP (mm)	MDP (mm)	Record length (years)
1	Izeh	764	31°49′	49°51′	603.7	5.5	32
2	Pole Shalu	700	31°45′	50°08′	762.8	6.7	30
3	Susan	600	31°59′	49°52′	766.6	6.1	31
4	Andika	500	33°02′	49°24′	547.7	6.2	22
5	Lali	150	32°17′	49°03′	425.1	3.6	26
6	Abasspoor	820	32°04′	49°36′	552	5.1	32
7	Gotvand	75	32°15′	48°49′	384.5	4.7	31
8	Arabhasan	33	31°51′	48°53′	269.2	3	28
9	Ahvaz	20	31°20′	48°41′	219.4	4	42
10	Tange Panj	540	32°56′	48°46′	1140	8.3	23
11	Sade Dez	525	32°33′	48°27′	476.2	5.3	31
12	Sade Tanzimi	142	32°25′	48°27′	362.9	3.9	28
13	Chamgaz	38	32°57′	47°49′	481.3	4.8	26
14	Paye pol	90	32°25′	48°09′	293.9	4.3	16
15	Abdolkhan	40	31°50′	48°23′	226.9	3.7	31
16	Bagh Malek	675	31°33′	49°52′	563.9	6.4	39
17	Idanak	560	30°57′	50°25′	617	8.3	29
18	Machin	380	31°23′	49°43′	372.2	4.9	25
19	Sade Shohada	333	30°40′	50°17′	340.8	3.8	42
20	Kamp Jarahi	8	30°43′	49°11′	187.6	3.1	24
21	Dehmolla	32	30°30′	49°40′	220.6	3.7	31

Kyselý et al. (2007), Jingyi and Hall (2004), and Kjeldsen et al. (2002) studies in the application of statistical methods have confirmed L-moments and showed that PWMs and L-moments are preferred to the classic estimation methods, especially for the regional studies.

Using L-moments and regionalizing the GEV distribution, Ariff et al. (2016) extracted the regional IDFs for the Malaysian Peninsular region. They evaluated the application of this method due to the associated simplicity as well as its efficiency in areas without appropriate stations.

Eslamian and Feizi (2007), using L-moments and GEV distribution, compared the regional frequency analysis of maximum monthly rainfall in the Isfahan region, Iran, with their single-site values. They described the L-moments method as an accurate and helpful tool for confirming the similarities or differences in regional rainfall frequency analysis.

2 Study area and data

This study provided data from 21 rain gauge stations located in Khuzestan province, Iran, by Province Water and Electricity Organization. The minimum record length related to the Payepol station was 16 years, and the maximum record length related to Ahwaz stations and the Shohada dam was 42 years (Table 1). The data used to determine the number of optimal clusters and also the input data to clustering models include the geographical latitude and longitude variables, height from sea level, maximum average precipitation (MAP), maximum daily precipitation (MDP), annual rainfall average for each station which has been shown in Table 1. Khuzestan province, which covers 4% of the country's total area, is the largest in the western half of the country. This province is located between 47°41′ to 50°39′ east longitude and 29°58′ to 33°04′ north of the equator. Despite having only 4% of the country's total area, this province owns more than 30% of the country's surface water.

3 Methodology

In this section, a suggested method to extract the IDF regional curves has been explained; the steps are as follows:

1- Determine the number of optimal clusters using hydrologic data provided in Table 1.
2- Implement of neural gas networks and other common models mentioned in the clustering.
3- Investigate the homogeneity of the formed regions and the discordance of the stations in each region.

4- Determine the appropriate statistical distribution for each region and estimate the distribution parameters at the required duration.

5- Investigate the quantiles or amounts of precipitation in duration and the required return periods.

6- Draw the regional IDF curves.

3.1 Neural gas (NG) network

The rule of learning in the neural gas network is as follows:

$$w_i^{new} = w_i^{old} + \alpha_i(x - w_i^{old}) \tag{1}$$

$$\alpha_i = \varepsilon e^{\frac{-k_i}{\lambda}} \tag{2}$$

where w_i is a gas molecule formed in data space. The number of these molecules is initially assumed as a value, and eventually, it is revised to have the logical and optimal function of the algorithm. These elements also have been selected in the main data range. α_i is a parameter that specifies the learning rate and depends on k_i and λ. As if λ tends to infinity, learning of the whole neurons would be equal and if it tends to zero, then the nearest neuron begins to learn. The extreme modes of λ are not suitable alone, and usually, a mode between them is chosen. k_i refers to the superior neuron to the i neuron. $\mathcal{E}$ is also a constant number that controls the learning rate.

To create a neighborhood between the first and second neurons in terms of proximity, an edge is created. For each neuron, there is $c_{i.j}\epsilon\{0.1\}$ which shows that there exists an edge or neighborhood or does not exist and also $t_{i.j}\epsilon\{0.1.2\ldots.\}$ which shows the time intervals (age) from the last meeting or re-edge, that if it exceeds more than one size, the neighborhood will be broken. This approach helps the neural network to learn topology.

NG algorithm can be summarized as follows:

Step 1: A random position of w_i is created in the data space.
Step 2: An input named x is selected from the expected data.
Step 3: Aging, which includes computation of the distance between x, and the centers of w_i and k_i aging for each center.
Step 4: Adaption or learning.

$$w_i^{new} = w_i^{old} + \varepsilon e^{\frac{-k_i}{\lambda}}(x - w_i^{old}) \tag{3}$$

The main point is that during the training period, as the algorithm progresses, the learning speed should be reduced; otherwise, the neural network will be repeated, and an incorrect cycle will be created. For this purpose, the amount of λ and ε should be decreased as learning progresses. So, the following function would be used.

$$\begin{aligned} &\mathrm{G(t)} = G_i\left(\frac{G_f}{G_i}\right)^{\frac{t}{t_{max}}} \\ &\lambda_i > \lambda_f, \varepsilon_i > \varepsilon_f, T_i < T_f \end{aligned} \tag{4}$$

where i index shows the parameter value at the beginning of learning and f index shows the value of the parameter at the end of the learning process. For instance, if $t=0$, so $\lambda(t) = \lambda_i$, and if$t = t_{\max}$, so$\lambda(t) = \lambda_f$.

Step 5: An edge between the first two ranks in terms of proximity and age of this edge is considered equal to zero (create a neighborhood).
Step 6: Age of all edges increases ($t_{i.j} \rightarrow t_{i.j} + 1$)
Step 7: It is assumed that $k_i = 0$, and for each j which is $t_{i.j} > T$, it is considered as $c_{i.j} = 0$. At this step for the reasons mentioned in step 4, T should be increased during the learning period to reduce the degree of rigidity, which means the edges are allowed to last longer.
Step 8: If the termination conditions are not met (for example, the maximum quantity of neurons or any amount of performance), the step 2 is repeated. Otherwise, algorithm steps would be finished.

3.2 Growing neural gas (GNG) network

GNG algorithm, which is based on unsupervised artificial neural networks, was first introduced by Fritzke (1994). The GNG network is a clustering algorithm working step by step; the number of neurons increases without using previous knowledge about the structure of input patterns during the learning process (Fink et al. 2015). Unlike classical clustering algorithms, the GNG algorithm owns a compatible network structure which makes it suitable for learning the large data set topologies (Zaki and Yin 2008). The main idea of GNG is that it will continuously add the new nodes (neurons) to a small initial network in a growing structure. In the GNG network, the neurons compete to determine which one is most similar to the input data set (Morell et al. 2014).

GNG algorithm can be summarized as given below:

Step 1: Creating two random neurons at locations w_1 and w_2
Step 2: Selecting vector input called x
Step 3: Finding the best neuron (s_1) and second-best neuron (s_2)
Step 4: Increasing age of all edges connected to s_1

$$\forall_j : t_{s_1 j} \leftarrow t_{s_1 j} + 1$$

Step 5: Increasing the amount of accumulated error in s_1

$$E_{s_1} = E_{s_1} + \Delta E_{s_1} \tag{5}$$

$$\Delta E_{s_1} = \|w_{s_1} - x\|^2 \tag{6}$$

Step 6: Adaptation

$$w_{s_1}^{new} = w_{s_1}^{old} + \varepsilon_b\left(x - w_{s_1}^{old}\right) \tag{7}$$

$$\begin{gathered} w_n^{new} = w_n^{old} + \varepsilon_n\left(x - w_n^{old}\right) \\ \varepsilon_b > \varepsilon_n \end{gathered} \tag{8}$$

Step 7: Creating an edge between s_1 and s_2 if there is not any.

$$C_{s_1 s_2} = 1.t_{s_1 s_2} = 0$$

Step 8: All edges that their age is more than T will be deleted.

$$t_{ij} > T \rightarrow C_{ij} = 0$$

Step 9: If the number of inputs presented to the network is an integer multiplier of L, a new neuron is created. This neuron is created at the location of w_r.

$$\begin{gathered} w_r = \tfrac{1}{2}\left(w_q + w_f\right) \\ C_{fq} = 0.\ C_{rf} = C_{rq} = 1 \end{gathered} \tag{9}$$

q is the neural index which has the most amount of accumulated error; f is the neighbor index of q which has the most errors.

E_f and E_q errors with α coefficient are declined:

$$E_f \leftarrow \alpha E_f \quad . \quad E_q \leftarrow \alpha E_q \quad \alpha < 1.$$

Consider error E_r equals to E_q. $E_r = E_q$

Step 10: Decreasing the accumulated error of all neurons. $E_i \leftarrow dE_i d < 1$
Step 11: If the stop measurement (for example maximum number of neurons or any scale of performance) has not yet been met, step 1 would be repeated.

3.3 Discordancy and heterogeneity measures

An area containing N stations is considered so that the i station has the record length of n_i and the ratio of L-moments $t^{(i)}$. $t_3^{(i)}$ and $t_4^{(i)}$. In this case, the discordancy criterion D_i would be calculated using the below relations.

$$u_i = \left[t^{(i)}.t_3^{(i)}.t_4^{(i)}\right]^T \tag{10}$$

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i \tag{11}$$

$$s = (N-1)^{-1}\sum_{i=1}^{N}\left(u_i - \bar{u}\right)\left(u_i - \bar{u}\right)^T \tag{12}$$

$$D_i = \frac{1}{3}\left(u_i - \bar{u}\right)^T\left(u_i - \bar{u}\right)S^{-1} \tag{13}$$

where $u_i = \left[t^{(i)}.t_3^{(i)}.t_4^{(i)}\right]^T$ is the L-moment ratio matrix in station i, N is the number of stations, and S is the sample covariance matrix.

If D_i is big, the location i is discordant. An appropriate criterion to determine if a station is discordant or not is that D_i is bigger than 3 or equal to it.

To calculate the degree of heterogeneity, first V_1 would be obtained using Eq. (14) for the observed data.

$$V_1 = \sum_{i=1}^{N} n_i\left(t^{(i)} - \bar{t}\right)^2 / \sum_{i=1}^{N} n_i \tag{14}$$

$$\bar{t} = \left(\sum_{i=1}^{N} N_i t^{(i)}\right) / \left(\sum_{i=1}^{N} n_i\right) \tag{15}$$

where n_i is the size of samples in the station i, $t^{(i)}$ is the sample L-moment (L-CV), $\bar{t}$ is the point average of sample moment (L-CV).

For each simulated area, V_1 would be calculated. Also, from simulated data, average μ_v and standard deviation σ_v and inhomogeneity criterion would be determined through relation 16.

$$H_i = \frac{(V_i - \mu_v)}{\sigma_v} \tag{16}$$

Hosking and Wallis (1997) suggested that an area can be an acceptable homogenous area if H_i is smaller than 1, and it can be relatively heterogeneouss if H_i is between 1 and 2, and it would be definitely heterogeneous if H_i is bigger than 2. In practice, the H_1 criterion is more appropriate (Rao and Srinivas 2006).

3.4 Selecting the appropriate distribution

Selecting an appropriate frequency distribution for homogeneous regions can be done by comparing the distribution moments with the average regional moment of the data. Also, to select the best distribution, a goodness of fit test will be performed for the distribution function. This test would be done through calculation statistics of Z^{Dist}. An appropriate distribution function is a function which is $\left|Z^{Dist}\right| < 1.64$.

$$Z^{Dist} = \left(\tau_4^{Dist} - \bar{\tau}_4 + \beta_4\right)/\sigma_4 \tag{17}$$

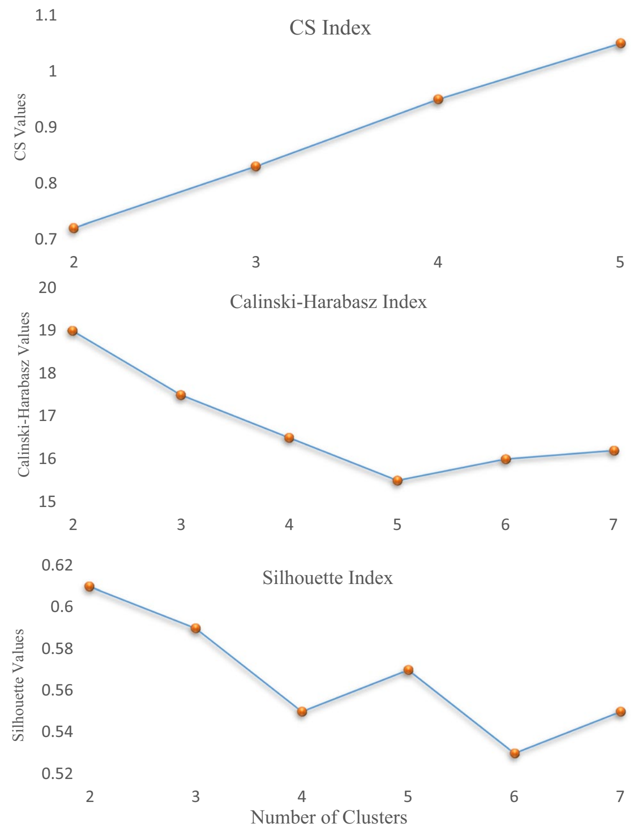

Fig. 1 CS, Calinsky-Harabasz, and Sillhouette values for determining the optimal number of clusters

$$\beta_4 = N_{sim}^{-1}\sum\nolimits_{m=1}^{N_{sim}}\left(\overline{\tau}_{4m} - \overline{\tau}_4\right) \tag{18}$$

$$\sigma_4 = \left\{ \left(N_{sim} - 1\right)^{-1} \sum_{m=1}^{N_{sim}} \left(\overline{\tau}_{4m} - \overline{\tau}_4\right)^2 - N_{sim}\beta_4^2 \right\}^{1/2} \tag{19}$$

Here "dist" means distribution, τ_4^{Dist} is the size or distribution kurtosis criterion (LC_K), $\overline{\tau}_4$ is the areal average of L-moment sample kurtosis, β_4 the area bias value of the above moment, σ_4 is the regional deviation of the above moment, and N_{sim} is the number of simulated areas.

3.5 Regionalization of IDF curves

A set of desired quantiles for the station j which has a record length of n_j and located in a region with the N_s stations are shown by Q_j. Rainfall observation data, x_j for the station in specified quantiles, T would be calculable through Eq. (20). So, rainfall data sets in the station j can be calculated as follows:

$$\begin{gathered} x_{j.k} = Q_j\left(T_k\right) \\ k = 1 \dots .n_j \;\; . \;\; j = 1 \dots .N_s \end{gathered} \tag{20}$$

If the area is homogeneous, the set of quantiles for the station j will be as per Eq. (21).

$$T = 2, 5, 10, 20, 50, 100j = 1, \dots, nQ_j(T) = \mu_j X_T \tag{21}$$

In Eq. (21), X_T is a set of dimensionless regional quantile with a probability of not exceeding f which is called the regional growth curve. μ_j is the scale factor for station i, where parameters such as mean or median are considered to simplify the calculations.

The value of variation coefficient moment and ratios of L-moments for the station j using single site data, x_j, is equal to their amounts for regional data. As a result, it will be possible to estimate the regional quantiles X, by equating the first to fourth moments of the region with the mean, the coefficient of variation moments, and the L-moments ratios of the distribution function considered for the region.

By estimating the set of quantiles X, for the maximum annual series of rainfall intensities in each duration and the desired return period in a homogeneous region, along with estimating the scale factor μ_j, for only one station in the region, different values (i, d, T) using Eqs. (20) and (21) will be computable. So, it is not needed to estimate the probability of distribution function for every single annual series in each station. Finally, using these values, a regional IDF curve will be drawn for each homogeneous region.

To investigate the differences between the regional IDF curves which are based on the regional distribution functions with the stationary IDF curves, three equations of the coefficient of variation of root mean square error (CV_{RMSE}), mean percentage difference (Δ), and mean bias error (MBE) as per below were used. The lower the CV_{RMSE} and Δ values, the more accurate the model used in clustering. Also, the negative MBE values indicate overestimation, and the positive values indicate an underestimation of the regional values than the at-station rainfall values.

$$CV_{RMSE} = \frac{\sqrt{\frac{1}{N_d N_T}\sum_{d=1}^{N_d}\sum_{T=1}^{N_T}\left(x_{d.T} - z_{d.T}\right)^2}}{\frac{1}{N_d N_T}\sum_{d=1}^{N_d}\sum_{T=1}^{N_T} x_{d.T}} \times 100 \tag{22}$$

$$\Delta = \frac{1}{N_d N_T}\sum\nolimits_{d=1}^{N_d}\sum\nolimits_{T=1}^{N_T}\frac{\left|x_{d.T} - z_{d.T}\right|}{x_{d.T}} \times 100 \tag{23}$$

$$MBE = \frac{\sum_{d=1}^{N_d}\sum_{T=1}^{N_T}\left(x_{d.T} - z_{d.T}\right)}{N_d N_T} \tag{24}$$

In the above relations, $x_{d.T}$ and $z_{d.T}$ are respectively maximum rainfall intensity in duration d and the return period T

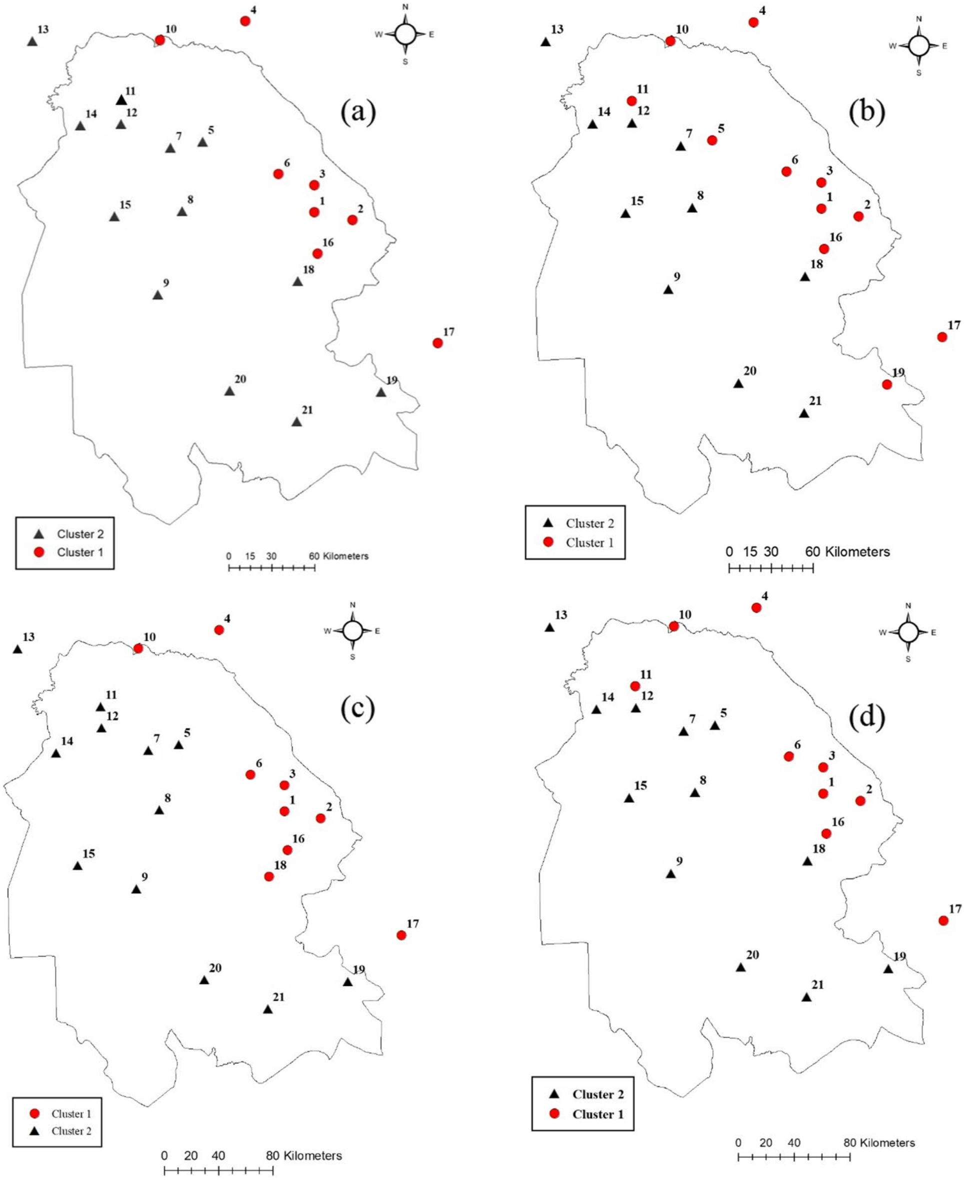

Fig. 2 The result of the location of stations in the clusters was identified by (**a**) GNG method, (**b**) NG method, (**c**) FCM method, and (**d**) SOM, K-means, and Ward methods

in the specified station and the homogeneous area in which the station is located. Also, N_d and N_d are the number of durations and return periods.

4 Results

4.1 Cluster analysis

Used data sets should be normalized before entering the clustering models. This is due to data from the different types, such as geographical and precipitation data, which also have different units. Based on these normalized data, probabilistic homogeneous regions were determined using clustering models, including the new method of neural gas networks and the common models of Ward, K-means, self-organizing map, and fuzzy C-means.

CS (Chou et al. 2004), Silhouette (Rousseeuw 1987), and Calinski-Harabasz (Caliński and Harabasz 1974) indices were used to determine the optimal number of clusters. Figure 1 shows the number of optimal clusters in a range of clusters. Since the highest value in Silhouette and

Table 2 Employed parameters in the NG and GNG

Model	Parameters
Ng	$T_i = 5T_f = 1t_{\max} = 10,000\ \varepsilon_i = 0.9\varepsilon_f = 0.001\ \lambda_i = 1\ \lambda_f = 0.5$
GNG	$\mathrm{T} = 50\ L = 40\ \ d = 0.995\ \ \ \alpha = 0.5\ \ \ \varepsilon_{\mathrm{n}} = 0.0006\ \varepsilon_{\mathrm{b}} = 0.05$

Table 3 Results of heterogeneity and discordancy measures for 24-h rainfall duration

Cluster	Clustering models	Number of stations	Discordant stations	Heterogeneity measure			Heterogeneity situation
				H_1	H_2	H_3	
1	NG	11	-	−1.12	−1.13	−1.04	Definitely homogeneous
	GNG	8	-	−0.91	−1.91	−1.63	Definitely homogeneous
	SOM	9	-	−1.11	−2.01	−1.81	Definitely homogeneous
	FCM	9	-	−0.64	−1.79	−1.59	Definitely homogeneous
	K-means	9	-	−1.11	−2.01	−1.81	Definitely homogeneous
	Ward	9	-	−1.11	−2.01	−1.81	Definitely homogeneous
2	NG	10	-	0.17	−1.28	−2.33	Definitely homogeneous
	GNG	13	-	0.23	−0.12	−1.3	Definitely homogeneous
	SOM	12	-	0.39	−0.04	−1.21	Definitely homogeneous
	FCM	12	-	0.12	−0.01	−1.09	Definitely homogeneous
	K-means	12	-	0.39	−0.04	−1.21	Definitely homogeneous
	Ward	12	-	0.39	−0.04	−1.21	Definitely homogeneous

Table 4 Estimated parameters of GEV and GLOG distributions as the regional probability distributions of annual maximum storm intensities for two regions in the NG model

Cluster 2 (GEV)			Cluster 1 (GLOG)			*d* (min)
U	*α*	*K*	*ε*	*α*	*K*	
0.7542	0.4257	−0.0002	0.8744	0.2624	−0.2672	15
0.7607	0.4258	0.0115	0.8836	0.2674	−0.2464	30
0.7571	0.4123	−0.0119	0.8792	0.2582	−0.2621	45
0.7574	0.4040	−0.0230	0.8808	0.2547	−0.2621	60
0.7552	0.3892	−0.0499	0.8810	0.2489	−0.2669	90
0.7568	0.3762	−0.0657	0.8867	0.2451	−0.2595	120
0.7544	0.3728	−0.0766	0.9077	0.2366	−0.2235	180
0.7622	0.3930	−0.0276	0.9426	0.2334	−0.1458	360
0.7635	0.4183	0.0121	0.9721	0.2402	−0.0701	720
0.7548	0.4243	−0.0007	09,646	0.2553	−0.0837	1080
0.7756	0.4117	0.0337	0.9621	0.2533	−0.0901	1440

Calinski-Harabasz indices and the lowest value in CS indicate the number of optimal clusters, the number of region in all three models is equal to 2.

The output of clustering models based on two separate regions in the specified area, which demonstrates how stations are divided between these two regions, is illustrated in Fig. 2, as it is known that the result of clustering for some stations is different in several models which is due to the differences in the performance of each model. The Ward, SOM, and K-means models have shown the same performance. Also, region 1 occupies more eastern areas, and region 2 occupies the central and western areas of the province. According to the height index of each station, the results show that the stations with the higher altitudes in the eastern cluster and the lower stations in the western cluster are divided, which confirms the proper functioning of neural gas networks in terms of topographic detection of the data space.

The constant parameters used in both neural gas and growing neural gas networks are presented in Table 2.

Table 5 Average values of goodness-of-fit indices of the difference between IDF curves based on regional and at-site probability distributions for the used clustering models

Model	Values of goodness of fit indices		
	CV_{RMSE}	Δ	MBE
NG	24.39	15.56	−0.04
GNG	24.41	16.03	0.23
FCM	25.19	16.35	0.32
SOM, K-means, Ward	24.67	15.91	0.31

4.2 Regional homogeneity tests

The H and Di statistics, which are tests based on L-moments, were used to investigate the regional homogeneity and

Table 6 Values for goodness-of-fit indices of the difference between IDF curves based on regional and at-site probability distributions at 21 rainfall stations for the NG clustering model

Stations	Values of goodness-of-fit indices		
	CV_{RMSE}	MBE	Δ
1	29.46	− 2.11	10.75
2	42.46	− 6.61	24.84
3	26.38	− 2.89	11.67
4	14.68	− 1.42	10.04
5	15.61	− 1.77	9.14
6	23.07	− 2.05	8.51
7	17.49	0.28	14.29
8	23.38	− 0.11	10.32
9	13.59	1.43	5.91
10	21.3	6.85	28.14
11	18.67	1.58	8.33
12	16.27	0.80	13.06
13	23.36	− 1.38	20.99
14	46.52	9.48	28.56
15	27.05	− 2.34	14.41
16	26.02	− 3.71	18.87
17	29.19	0.27	19.62
18	15.55	2.13	22.00
19	35.81	− 3.70	21.42
20	30.17	5.12	14.33
21	16.11	− 0.69	11.72
Average	**24.39**	**− 0.04**	**15.56**

discord of the stations in each region. These values were determined for the maximum intensity of annual rainfall at different duration, as well as the various models used for clustering. These results are presented for 24-h rainfall in Table 3. Referring to the results, in none of the applied clustering models in different time durations, there was no discordant station. Except for a few cases, all of the regions formed in different models were homogeneous, which indicates the reasonable accuracy of the clustering models used.

According to the goodness of fit test results, the generalized logistics distribution (GLOG) and GEV distribution are selected as the regional distribution function for regions 1 and 2, respectively. To estimate the required quantiles, it is necessary to calculate the parameters of regional distribution. For this purpose, in all of the used clustering models, the first to fourth moments of both generated regions are considered equal to the first to fourth moments of the distribution function considered for the region. The neural gas network model (NG) results are presented as an instance in Table 4.

To determine the best clustering model, the numerical values of regional curves with the same stationary curves were compared. As shown in Table 5, according to the calculated estimate error values, the neural gas network clustering model has the lowest error amount in both indicators, which shows the superiority of this method over the other methods. Also, the negative MBE index for this model indicates that the numerical values of the regional IDF curves obtained from this method are somewhat more significant than the at-station values (overestimation). Also, by considering all three indicators, the growing neural gas network can be considered the second suitable model. By considering the error values between the regional IDF with at-station IDF, which are presented in Table 6, it is concluded that in stations with a short record length, the estimated error amount has been increased, and if this station is removed, the better performance can be expected from the used clustering models. The comparison between the two types of regional and at-site curves in the four selected stations is illustrated in Fig. 3.

5 Conclusions

In this study, two new models of neural gas and growing neural gas networks were presented to regionalize the IDF curves. For this purpose, taking into account the characteristics of longitude, latitude, average annual rainfall, altitude, and maximum 24-h annual rainfall for each station, and using three indicators and CS, Silhouette, and Calinski-Harabasz(CH), it has been determined that Khuzestan province has two separate and possibly homogenous regions. Then, using different clustering models, homogeneous regions have been formed. Clustering was one of the most important and main steps of this research due to the associated sensitivity and great impact on the final result. Therefore, clustering operations were performed using six different models, Ward, K-means, FCM, and Self-organizing map (SOM), which are among the most widely used methods. In addition to the four methods mentioned, two new models of the neural gas network (NG) and growing neural gas network (GNG) were used for clustering.

To investigate the homogeneity of the two regions, as well as the discordancy of the stations in each region, in all the six models in eleven durations, the regional homogeneity tests and discordancy tests based on L-moments were used. In most models and different durations, the regions created by clustering had a good homogeneity. After determining the position of each station in the dual regions and identifying both areas as homogeneous, the regional distribution function was determined, and then the regional IDFs were extracted using the L-moments method. The regional IDF curve obtained for each area was compared with the at-station IDF curves in the same area. The results showed that in all of the stations, the regional and stationary

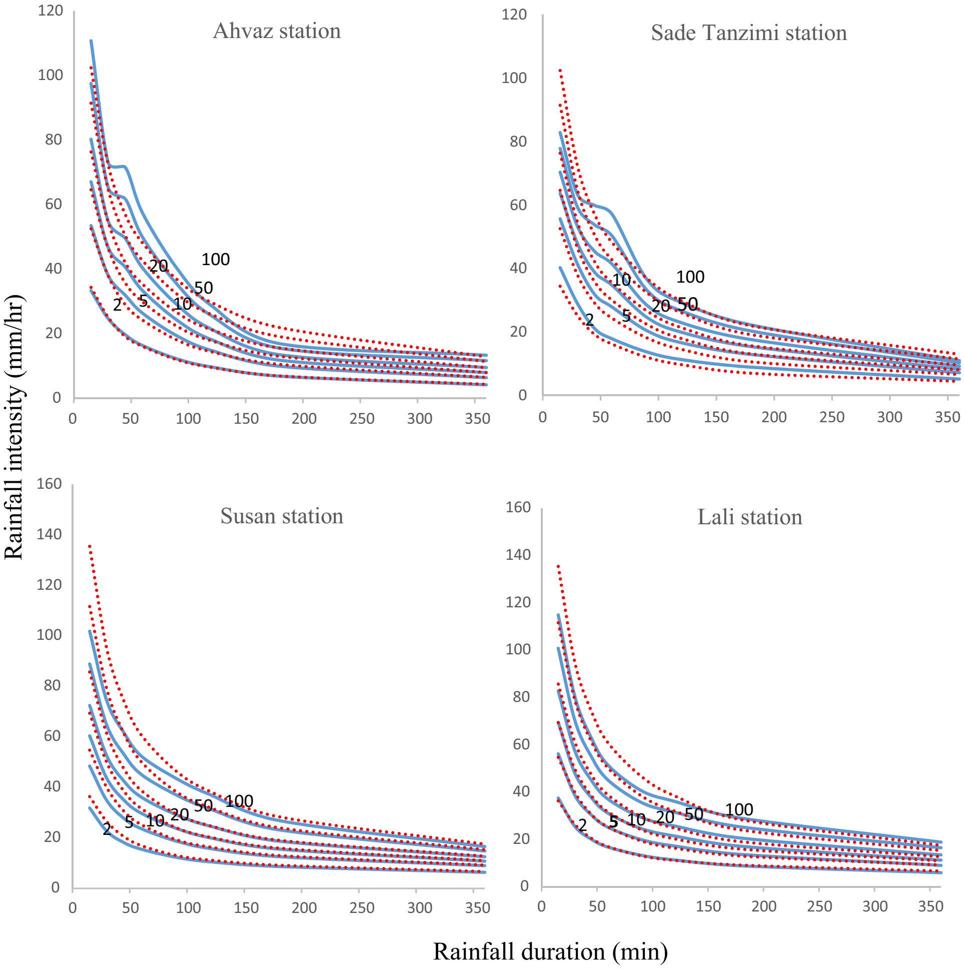

Fig. 3 IDF curves for four rainfall stations in Khuzestan province (Note: The values on each curve are the return period (T), blue lines are at-site, and red lines are regional IDF curves)

IDF curves are highly consistent and show the same trend. This research is the first one to evaluate the efficiency of neural gas networks in regionalizing the IDF curves. Among extracted regional IDFs, the curves obtained from the region composed of neural gas networks and growing neural gas network models had the highest accuracy and the most compliance with the at-station curves, which indicates the efficiency of these models in terms of regionalization. The quality of operation of neural gas networks can improve the various issues and problems related to water resources management and planning.

Acknowledgements The authors appreciate the constructive comments of anonymous reviewers on this paper, which helped improve the final version of the paper.

Author contribution All authors collaborated in the research presented in this publication by making the following contributions: research conceptualization, M.R.M., S.E., S.S., and M.T.; methodology, M.R.M., S.E., and S.S.; formal analysis, M.R.M. and S.E.; writing—original draft preparation, M.R.M.; writing—review and editing, M.R.M, S.E., and M.T.; supervision, S.E.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Ethics approval We confirm that this article is an original research and has not been published or presented previously in any journal or conference in any language.

Consent to participate Not applicable.

Consent for publication All the authors consented to publish the paper.

Conflict of interest The authors declare no competing interests.

References

Abdi A, Hassanzadeh Y, Ouarda TB (2017) Regional frequency analysis using growing neural gas network. J Hydrol 550:92–102

Alemaw BF, Chaoka RT (2016) Regionalization of rainfall intensity-duration-frequency (IDF) curves in Botswana. J Water Resour Prot 8(12):1128

Amin MZM, Shaaban, AJ (2004) The rainfall intensity-duration-frequency (IDF) relationship for ungauged sites in peninsular

Malaysia using a mathematical formulation. In Proceedings 1st International Conference on Managing Rivers in the 21st Century, River Engineering and Urban Drainage Research Centre, Penang, MALAYSIA 251–258.

Angelopoulou A, Psarrou A, Garcia-Rodriguez J, Orts-Escolano S, Azorin-Lopez J, Revett K (2015) 3D reconstruction of medical images from slices automatically landmarked with growing neural models. Neurocomputing 150:16–25

Ariff NM, Jemain AA, Bakar MA (2016) Regionalization of IDF curves with L-moments for storm events. Int J Math Comput Sci 10:217–223

Bell FC (1969) Generalized rainfall-duration-frequency relationships. J Hyd Div 95:311–327

Bernard MM (1932) Formulas for rainfall intensities of long duration. Trans Am Soc Civil Eng 96(1):592–606

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Comm Stats-Theory and Methods 3(1):1–27

Carlevarino A, Martinotti R, Metta G, Sandini G (2000) An incremental growing neural network and its application to robot control. Proceeding of the International Joint Conference on Neural Networks, Como, Italy, Jul. 24-27, pp. 323-328

Chou CH, Su MC, Lai E (2004) A new cluster validity measure and its application to image compression. Pattern Anal Appl 7(2):205–220

Cselényi Z (2005) Mapping the dimensionality, density and topology of data: the growing adaptive neural gas. Comput Meth Programs Biomed 78:141–156

Decker R (2005) Market basket analysis by means of a growing neural network. Int Rev Retail Distrib Consum Res 15(2):151–169

Eslamian SS, Feizi H (2007) Maximum monthly rainfall analysis using L-moments for an arid region in Isfahan province. Iran J Appl Meteorol Climatol 46(4):494–503

Ferrer G J (2014) Creating visual reactive robot behaviors using growing neural gas. In Proceedings of the 25th Modern Artificial Intelligence and Cognitive Science Conference, Spokane, USA, Apr. 26, pp. 39–44.

Fink O, Zio E, Weidmann U (2015) Novelty detection by multivariate kernel density estimation and growing neural gas algorithm. Mech Syst Signal Proc 50:427–436

Fritzke B (1994) A growing neural gas network learns topologies. Advances in neural information processing systems. Proceedings of the 8th International Conference on Neural Information Processing Systems. MIT Press, Cambridge, MA

Greenwood JA, Landwehr JM, Matalas NC, Wallis JR (1979) Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour Res 15:1049–1054

Hosking JR (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. J Roy Stat Soc: Ser B (methodol) 52(1):105–124

Hosking JRM, Wallis JR (1997) Regional frequency analysis. Cambridge University Press, Cambridge, p 240

Jingyi Z, Hall M (2004) Regional flood frequency analysis for the Gan-Ming River basin in China. J Hydrol 296:98–117

Kjeldsen TR, Smithers J, Schulze R (2002) Regional flood frequency analysis in the KwaZulu-Natal province, South Africa, using the index-flood method. J Hydrol 255:194–211

Koutsoyiannis D, Kozonis D, Manetas A (1998) A mathematical framework for studying rainfall intensity-duration-frequency relationships. J Hydrol 206:118–135

Kyselý J, Picek J, Huth R (2007) Formation of homogeneous regions for regional frequency analysis of extreme precipitation events in the Czech Republic. Stud Geophys Geod 51:327–344

Lee SH, Maeng SJ (2003) Frequency analysis of extreme rainfall using L-moment. Irrig Drain: J Int Comm Irrig Drain 52(3):219–230

Lisboa PJ, Edisbury B, Vellido A (2000) Business applications of neural networks: the state-of-the-art of real-world applications. World Scientific Publishing Company, Singapore

Martinetz T, Schulten K (1991) A "neural-gas" network learns topologies. Artificial Neural Network 1:397–402

Morell V, Cazorla M, Orts-Escolano S, Garcia-Rodriguez J (2014) 3d maps representation using GNG. In 2014 International Joint Conference on Neural Networks (IJCNN) Jul 6, 2014 (pp. 1482–1487). IEEE.

Oliveira Martins L, Silva AC, De Paiva AC, Gattass M (2009) Detection of breast masses in mammogram images using growing neural gas algorithm and Ripley's k function. J Signal Process Syst 55(1):77–90

Perica S, Martin D, Pavlovic S, Roy I, St Laurent M, Trypaluk C, Unruh D, Yekta M and Bonnin G (2013) Precipitation-frequency atlas of the United States, vol 9, version 2.0. Southeastern States; Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi

Rao AR, Hamed KH (1997) Regional frequency analysis of Wabash River flood data by L-moments. J Hydrol Eng 2:169–179

Rao AR, Srinivas V (2006) Regionalization of watersheds by hybrid-cluster analysis. J Hydrol 318:37–56

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Soltani S, Helfi R, Almasi P, Modarres R (2017) Regionalization of rainfall intensity-duration-frequency using a simple scaling model. Water Resour Manag 13:4253–4273

Yang T, Shao Q, Hao ZC, Chen X, Zhang Z, Xu CY, Sun L (2010) Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. J Hydrol 380:386–405

Zaki SM, Yin H (2008) A semi-supervised learning algorithm for growing neural gas in face recognition. J Math Model Algorithms 7:425–435