

## Bivariate regional drought frequency analysis using multivariate approaches: a case study in southwestern Iran

Hanie Pashaie\*, S. Saeid Mousavi Nadoushani, Ali Moridi and Ali Ahani

Department of Water Resources Management, Faculty of Civil, Water and Environmental Engineering, Shahid Beheshti University, Tehran, Iran

\*Corresponding author. E-mail: haniepashaie@gmail.com; h.pashaie@mail.kntu.ac.ir

### ABSTRACT

Bivariate approaches in Regional Frequency Analysis (RFA) address two issues: first, to evaluate the homogeneity of regions, and second, to estimate the joint return periods. This study was conducted to investigate the joint return period of a severe historical drought in southwestern Iran. Fifty-nine rain gauges were first clustered into three, four, and five regions using the fuzzy c-means clustering (FCM) algorithm. Then bivariate discordancy and homogeneity tests were applied to adjust the initial clusters. The results showed that only in the case of three clusters were all the regions homogeneous. Therefore, it can be inferred that combining clustering analysis and discordancy test is insufficient to form homogeneous regions. Finally, the joint return period, by choosing Generalized Logistic and Wakeby as marginal distributions and Clayton as a copula, was estimated for all the sites in the three regions. Since no three-parameter distribution function fitted well to the variable severity, the bivariate homogeneity index does not necessarily attest to region homogeneity regarding the marginal distribution functions. It is also deduced that sites with higher mean annual precipitation (MAP) and, correspondingly, higher elevation are more likely to experience shorter return periods of same drought events, in contrast to sites with lower MAP or lower elevation.

**Key words:** copula, clustering analysis, drought, homogeneity test, L-moments, regionalization

### HIGHLIGHTS

- The estimation of return periods for extreme events relies on the availability of sufficient data.
- Generalized Logistic and Wakeby are chosen as marginal distributions, and Clayton is selected as a copula.
- Sites with higher mean annual precipitation (MAP) and higher elevation are found to be more likely to experience shorter return periods of the same drought events.

## 1. INTRODUCTION

Frequency analysis is a valuable tool for estimating quantiles of extreme values, such as drought, flood, and storm events. Needless to say, reliable frequency analysis is highly dependent on data availability. Since data records are often short, the method called Regional Frequency Analysis (RFA) was developed to increase the reliability of the quantile estimates (Hosking & Wallis 1997). Additionally, RFA can be used to provide reliable analysis in ungauged sites (Sadri & Burn 2011).

To discuss exactly what RFA does, let's assume a data set that includes  $N$  sites, each containing an  $n$ -year data record. Thus, using regional data (with  $n \times N$  data) instead of individual site data (with a record length equal to  $n$ ) helps to retain longer data records and improve the reliability of estimates of at-site quantiles in frequency analysis. Therefore, RFA consists of two main steps: first, regionalization of the area, which involves identifying homogeneous regions in which the sites have the same frequency distribution; and second, estimating quantiles using an appropriate regional distribution function.

Various methods have been developed for regionalization, with cluster analysis being a standard statistical multivariate analysis technique used to divide data points into multiple groups. This method has been successfully applied to RFA (Raziei *et al.* 2008; Yoo *et al.* 2012; Shamshirband *et al.* 2015; Azam *et al.* 2018; Ahani *et al.* 2020; Rahman & Rahman 2020; Ahani *et al.* 2022).

It should be noted that sites in a cluster may not necessarily follow the same frequency distribution. To address this issue, Hosking & Wallis (1993) investigated the practical utility of RFA using the index-flood procedure by developing the application of L-moments in hydrology. They presented two tests based on L-moments to measure discordancy to identify

discordant sites and homogeneity in a region. Ngongondo *et al.* (2011) used the L-moments approach for RFA in southern Malawi. Using the concepts of entropy and L-moments to perform the Hosking–Wallis homogeneity test. Rajsekhar *et al.* (2013) extracted eight and nine homogeneous regions in Texas in terms of severity and duration of drought, apart. Using the Standard Precipitation Index (SPI) and Hierarchical Cluster Analysis, Gocic & Trajkovic (2013) divided Serbia into three regions and used discordancy and homogeneity tests to examine the homogeneity of the regions. With the help of FCM, Goyal & Sharma (2016) investigated drought in three states of western India; after calculating the SPI on different timescales, they optimized the number of clusters using five cluster validation indices; finally, the univariate Hosking–Wallis homogeneity test was applied to the clusters. Ahani *et al.* (2022) proposed a ranking method applied to Karun-e-bozorg basin in the southwest of Iran to identify a suitable regionalization for regional flood frequency analysis; the method was developed to compare and rank various regionalizations and choose the most appropriate one. Their method has an acceptable performance in identifying a regionalization that provides suitable homogeneity, quantile estimation accuracy, and average region size. Parvizi *et al.* (2022) implemented RFA of drought severity and duration in the Karkheh River Basin, Iran. They used the L-moments to investigate the delay of hydrological drought after the meteorological drought.

It should be noted that drought is a multivariate phenomena. Drought severity and duration are the two key variables that characterize droughts (Shiau & Shen 2001). Their multivariate nature adds to the complexity of drought regional frequency analysis. When variables are correlated, univariate approaches become impractical (Abbasian *et al.* 2021).

The requirement for multivariate approaches has led to the creation of multivariate homogeneity and discordancy tests by Chebana & Ouarda (2007), which were further developed with the introduction of L-comoments by Serfling & Xiao (2007). Applying these bivariate tests to regional flood data for Quebec, Canada, Chabana *et al.* (2009) concluded that the bivariate tests are appropriate and more efficient than univariate tests. Based on the measures developed by Chebana & Ouarda (2007), Sadri & Burn (2011) compared the performance of univariate and bivariate RFA; the results signified that bivariate RFA of drought outperforms univariate homogeneity and discordancy tests for adjusting the hydrological regions. Zhang *et al.* (2015) used the fuzzy c-means clustering (FCM) method and L-comoments to divide China into five homogeneous regions, and they adopted RFA. Abdi *et al.* (2017) applied univariate and bivariate homogeneity tests based on univariate and bivariate L-moments to check the homogeneity of their study area, whose results indicated that the bivariate method provided higher accuracy than the univariate one. Azam *et al.* (2018) performed hierarchical clustering and PCA and used bivariate homogeneity tests to confirm discordancy and homogeneity in Korean regions; the results of univariate and bivariate homogeneity and discordancy tests showed significant differences due to the inability of univariate measures to consider the joint behavior of duration and severity. Therefore, bivariate approaches are necessary for a thorough analysis of drought characteristics. Mortuza *et al.* (2019), to assess future drought trends in Bangladesh, utilized copula-based bivariate regional frequency analysis due to the strong correlation between drought duration and severity. Their findings concluded that compared to the bivariate drought frequency analysis, the standard univariate frequency analysis leads to underestimation or overestimation. Karahacane *et al.* (2020), in order to enhance the design of more efficient hydraulic structures in northern Algeria, conducted complete multivariate flood frequency analysis using copula and multivariate homogeneity tests considering the flood characteristics – duration, volume, and peak flow.

To conduct RFA, data are pooled from homogeneous regions and analyzed using frequency analysis. Given the multivariate nature of drought, both its severity and duration must be considered in characterizing its impact. Therefore, calculating bivariate return periods is more effective than using only univariate return periods. However, fitting classical bivariate distribution functions is challenging which is rooted in the fact that the marginal distribution functions of drought severity and duration may not be identical (Shiau 2006). This limitation has led water resources experts to explore copula functions as an alternative approach (Sadri & Burn 2014).

Copulas, introduced by Sklar (1973) and developed by Nelsen (2007), connect multivariate probability distributions to their one-dimensional marginal distribution while considering the existence of dependence and correlation among the random variables (Shiau 2006). These functions are very flexible in studying the correlation structure of continuous random variables, and their efficiency in hydrology has been shown by many researchers (Sadri & Burn 2014; Yu *et al.* 2014; Montaseri *et al.* 2018; Nabaei *et al.* 2019; Abbasian *et al.* 2021). Copulas do not face the limitations of classical multivariate distribution functions and offer the possibility of separating the study of the correlation structure of variables from the study of margins (Shiau & Modarres 2009; Mirabbasi *et al.* 2012). A copula can be utilized to separately model the marginal distributions of severity and duration, enabling each variable to have its distribution. The copula function then links these individual distributions to

create a joint distribution, which captures any relationship between the variables. This can help insurers analyse frequency more accurately.

Iran is home to vast arid and semi-arid areas. In 1999–2002, Iran experienced one of the worst droughts in its history, according to UN-SPIDER<sup>1</sup>. The severe lack of precipitation resulted in a total loss of \$3.5 billion, the death of 800,000 live-stock, and the drying up of major reservoirs and inland lakes (Amoli *et al.* 2015). Given the devastating impact of this drought, it is important to understand its characteristics and estimate the return period of such extreme events. Moreover, a significant number of studies have shown that bivariate tests outperform univariate ones in terms of both regionalization and frequency analysis. Therefore, utilizing the tests is a viable option for investigating regional frequency analysis in studies. This study applied bivariate RFA in three sub-basins: Karkheh, Karun, and Jarahi-Zohra, to estimate the bivariate return period for the most severe drought in the last four decades. This study divided the meteorological stations into three, four, and five clusters using FCM. Then, the initial clusters were adjusted with the help of the bivariate discordancy test. After that, the homogeneity of the formed regions was examined by the bivariate homogeneity test based on which all three clusters were compared. Eventually, the copula extracted and analyzed the joint return periods of the most severe droughts in each region.

## 2. STUDY AREA AND DATA

Generally, two types of data are of interest to identify regions in regional hydrology studies: (1) at-site statistics based on measurements related to drought characteristics and (2) other site descriptors known as site characteristics. Site characteristics are physical properties of the site that can be identified even before any data are measured. These usually include the site's geographical location, elevation, and other physical attributes. Mean Annual Precipitation (MAP) can also be considered a site characteristic since it is not necessarily directly related to the probability distribution (Hosking & Wallis (1997)).

The study was conducted in three major basins of Iran: Karkheh, Karun, and JarahiZohre. Monthly precipitation record and geographic locations of 59 trendless meteorological stations, using the Mann–Kendall test, with at least 30 years records were obtained from the Iran Water Resources Management Company.<sup>2</sup> Figure 1 shows the location of the 59 stations within the study area.

## 3. METHODS

### 3.1. Standardized precipitation index

The SPI is the most appropriate index for meteorological drought investigation (Xu *et al.* 2015; Nabaei *et al.* 2019). According to McKee *et al.* (1993), to measure the SPI time series, the long-term precipitation records were first fitted to a gamma distribution and subsequently to a normal distribution. In this study, a 6-month SPI was conducted (Svoboda *et al.* 2012). However, different timescales do not affect regionalization results (Raziei *et al.* 2012; Portela *et al.* 2015). Therefore, assuming the truncation level is equal to 0 (Shiau 2006), duration (the number of months that a drought event lasts) and severity (a cumulative deficiency of SPI during a drought event) was extracted by the Run Theory (Yevjevich 1967).

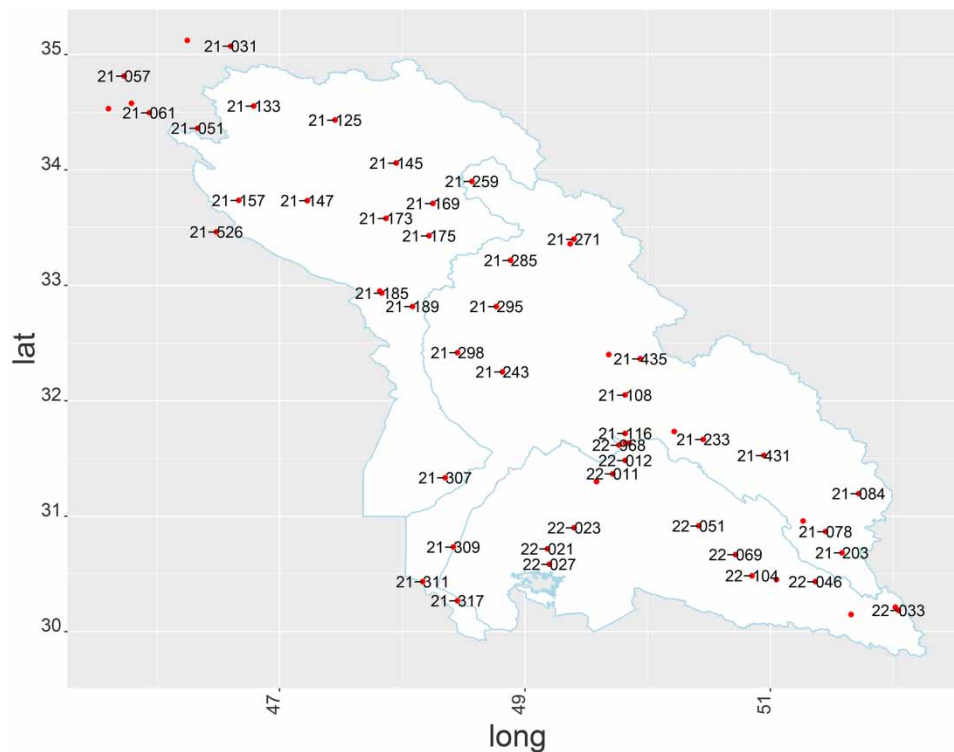
### 3.2. Formation of initial clusters

The purpose of cluster analysis is to divide a set of data into different groups so that members within a cluster are as similar as possible and members of different clusters are as dissimilar as possible (Rajsekhar *et al.* 2013). FCM developed by Dunn (1974) and Bezdek *et al.* (1984) is a soft clustering algorithm in which a data point belongs to every cluster with different degrees of membership (Sadri & Burn 2014; Zhang *et al.* 2015; Goyal & Sharma 2016). The objective function which has to be minimized in the FCM algorithm is as follows:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad (1)$$

<sup>1</sup> <http://www.un-spider.org>.

<sup>2</sup> <http://wrbs.wrm.ir>.



**Figure 1** | The location of the meteorological stations in the study area.

in which  $N$  is the number of data,  $C$  is the number of clusters,  $c_j$  denotes the center of the  $i$ th cluster,  $m$  is the fuzzifier, and  $u_{ij}$  is the membership degree of data  $j$  point in the cluster  $i$ .

In this study, longitude, latitude, and MAP were selected as inputs for initial clustering (Supplement, Table 6). It is noteworthy that the elevation of sites is usually correlated with MAP. The fuzziness of the FCM leads to different estimates for the same station, where by using the weighted average, a more reliable estimation of the bivariate return period can be provided for this category of stations.

In this study, initially, three clusters ( $nc = 3$ ) were created. Then, four ( $nc = 4$ ) and five ( $nc = 5$ ) clusters were added to investigate the possibility of having less discordant sites, which do not belong to any region. We did not pursue more clusters because increasing the number of clusters results in smaller region sizes and less data, which can lead to less reliable estimations.

### 3.3. L-moments

Developed by Hosking & Wallis (1993), the L-moment approach is an alternative system for describing probability distribution function. According to Hosking & Wallis (1997), L-moments have three important applications in RFA:

- Discordancy test: to identify discordant sites
- Homogeneity test: to assess if the region is homogeneous
- Fitting statistical distribution candidates and evaluating goodness-of-fit

L-moment ratios of order 1–4, including  $\lambda_1$ , L-CV ( $\tau$ ), L-skewness ratio ( $\tau_3$ ) and L-kurtosis ratio ( $\tau_4$ ) are mostly used to summarize distribution functions (Hosking & Wallis 1997). Multivariate L-moments, also called L-comoments,<sup>3</sup> for a random variable ( $X^{(j)}, j = 1, 2$ ) with distribution  $F_j$ , are matrices whose elements are calculated (Chebana & Ouarda 2007).

<sup>3</sup> 'co' indicates covariance and 'L' indicates linear.

### 3.4. Bivariate discordancy test

When it comes to the modification of the initial clusters, then, they are called 'regions'. The initial clustering is usually based on the *site characteristics*, while *at-site statistics* are preferred for subsequent testing of the homogeneity of a proposed set of regions (Sadri & Burn 2011). Following the formation of the initial regions, identifying discordant sites whose L-moments significantly deviate from the region's average L-moments is necessary to maintain the region's homogeneity (Hosking & Wallis 1997). The bivariate discordancy test proposed by Chebana & Ouarda (2007) for site  $i$  is written as  $D_i$ . For a region with  $N$  sites, the matrix  $U_i$ , defined as  $U_i = [\Lambda_2^{*(i)} \ \Lambda_3^{*(i)} \ \Lambda_4^{*(i)}]^t$ . The region's unweighted average is thus:

$$\bar{U} = N^{-1} \sum_{i=1}^N U_i \quad (2)$$

Assume that

$$S = \sum_{i=1}^N (U_i - \bar{U})(U_i - \bar{U})^t \quad (3)$$

The matrix  $D_i$  is defined as follows:

$$D_i = \frac{1}{12} N (U_i - \bar{U})^t S^{-1} (U_i - \bar{U}) \quad (4)$$

To transform the matrix from multidimensional space to the real line, it is possible to use a norm.

The critical value depends on the number of sites in the region. It is suggested that  $\|D_i\| \leq \frac{(N-1)}{3}$  exposes a discordant site.

In this study, the bivariate discordancy test was performed for each site in each cluster. It should be noted that removing a site changes the mean of the L-moments. Therefore, discordant sites were removed one by one until all sites in the group had a discordancy index value below the critical value. After all, the belongingness of the discordant sites to other clusters were checked.

### 3.5. Bivariate homogeneity test

Is the discordancy test enough to make a homogeneous region? In a homogeneous region, all sites have almost the same population L-moment ratios. Thus, the variation between a site of the region's L-moment ratios should not exceed a given value. It is reasonable to consider L-CV as between-site variation due to the fact that L-CV, compared to the variations in L-skewness or L-kurtosis, has a much more significant impact on the variance of the final estimates of all (Hosking & Wallis 1997). First, the weighted standard deviation of the at-site sample L-CV is as follows:

$$V = \left( \frac{\sum_{i=1}^N n_i (t^i - t^R)^2}{\sum_{i=1}^N n_i} \right)^{\frac{1}{2}} \quad (5)$$

in which  $N$  is the number of sites in the region,  $n_i$  is the record length of site  $i$ ,  $\tau$  is the L-moments and  $t^R = \frac{\sum_{i=1}^N n_i t^i}{\sum_{i=1}^N n_i}$ . The bivariate form of the statistic  $V$  is written as (Masselet *et al.* (2017)):

$$V_{\|\cdot\|} = \left( \frac{\sum_{i=1}^N n_i (\|\Lambda_2^{*(i)} - \bar{\Lambda}_2^*\|)^2}{\sum_{i=1}^N n_i} \right)^{\frac{1}{2}} \quad (6)$$

where  $\|\cdot\|$  is the norm of matrix  $V$ ; and  $\Lambda_2^{*(i)}$  is the L-covariation coefficient matrix for site  $i$  with record length  $n_i$ ,  $i = 1, \dots, N$ .

Then, Monte Carlo simulation method is used to calculate the given value, expected dispersion value. A large number of regions ( $N_{sim} = 1,000$ ) are simulated, all with the same number of sites and the same distribution. Kappa and Clayton, general and commonly used distributions, are used as marginal and joint distribution, respectively (Chebana & Ouarda 2007;



Sadri & Burn 2014). Parameter  $\theta$  for Clayton copula can be measured as  $\theta = 2 \times \bar{\tau} / (1 - \bar{\tau})$  in which  $\bar{\tau}$  is the regional average of Kendall rank correlation coefficient. The L-moments and L-comoments can be calculated from the generated data for any site. It is noteworthy that the sites have record lengths the same as those of the observed data. With the regional average L-comoment ratios for simulated region, it is possible to calculate  $V_{sim}$  (statistic  $V$  for simulated region) from the Equation (6). Then the average and the standard deviation of  $V_{sim}$  ( $\mu_v$ ,  $\sigma_v$ ) for all  $N_{sim}$  regions are calculated. Finally, the homogeneity index is computed (Masselet *et al.* 2017).

$$H_{||\cdot||} = \frac{V_{||\cdot||} - \mu_{V_{sim}}}{\sigma_{V_{sim}}} \quad (7)$$

The critical values for  $H$  are recommended as below:

$$\begin{cases} H \leq 1 & \text{homogeneous} \\ 1 < H \leq 2 & \text{probably homogeneous} \\ 2 < H & \text{non homogeneous} \end{cases} \quad (8)$$

### 3.6. Marginal distribution

Five widely used three-parameter probability distribution functions in hydrology, including (1) Generalized Logistic, (2) Generalized Extreme-value, (3) Generalized Normal, (4) Pearson type III, and (5) Generalized Pareto, were candidates for selecting the best marginal distribution functions for the two variables of drought severity and duration.

Regarding the calculation of the goodness of fit of the drought severity, the  $Z^{\text{Dist}}$  statistic at a significance level of 10% was chosen, which was previously used in studies of regional floods and droughts (Hosking & Wallis 1997; Sadri & Burn 2014). However, due to the limited variety of drought duration values, the Kolmogorov-Smirnov (KS) test was used in this study to calculate the goodness-of-fit of duration.

The parameters of the marginal distribution functions were estimated by L-moments.

### 3.7. Copula

Copulas are functions that connect multivariate probability distributions to their one-dimensional marginal probability distribution, while considering the correlation between random variables. Let  $X$  and  $Y$  be fitted by probability density functions  $f_X(x)$  and  $f_Y(y)$  respectively; then CDF of the variables,  $F_Y(y)$  and  $F_X(x)$ , are the variables with the univariate pdf  $U(0, 1)$  according to Sklar (1973). If  $F_X(x)$  and  $F_Y(y)$  are marginal distribution for  $F$  continuously, there exists a unique Copula which is a CDF and its margins are univariate.

It means  $C : [0, 1]^2 \rightarrow [0, 1]$  where:

$$F(X, Y) = C(F_X(x), F_Y(y)) \quad (9)$$

There are six families of copula, of which Archimedean copulas are the most famous. Clayton, Frank and Gumbel are most commonly used in hydrology (Shiau 2006; Sadri & Burn 2014; Nabaei *et al.* 2019; Mortuza *et al.* 2019; Karahacane *et al.* 2020), and in this study were candidates for joint probability distribution functions. The best copula was selected based on the maximum pseudo-likelihood. Also, the candidate parameters were estimated by maximum pseudo-likelihood.

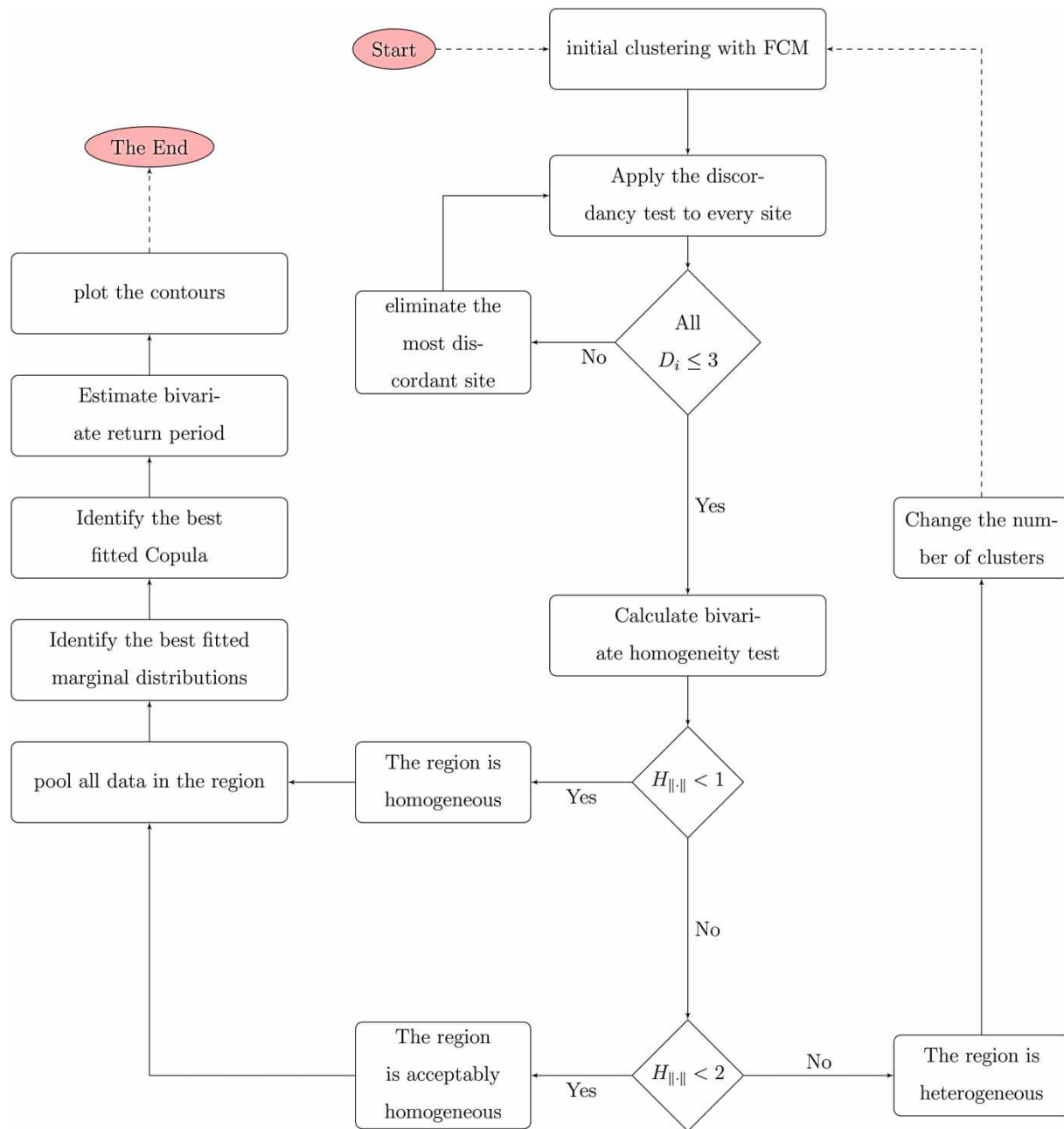
The steps of RFA are represented in the flowchart in Figure 2. The flowchart outlines the main steps for RFA in this study. All computations and statistical analyses and graphs in this study were performed using the 'R' software.

## 4. RESULTS AND DISCUSSION

### 4.1. Formation of initial clusters

The FCM was implemented to create the initial clustering with three standardized inputs: MAP, longitude, and latitude.  $\mathbf{a}_1$ ,  $\mathbf{b}_1$ , and  $\mathbf{c}_1$  in Figure 3 show the initial clusters using fcm for ( $nc = 3$ ), ( $nc = 4$ ), and ( $nc = 5$ ), respectively.

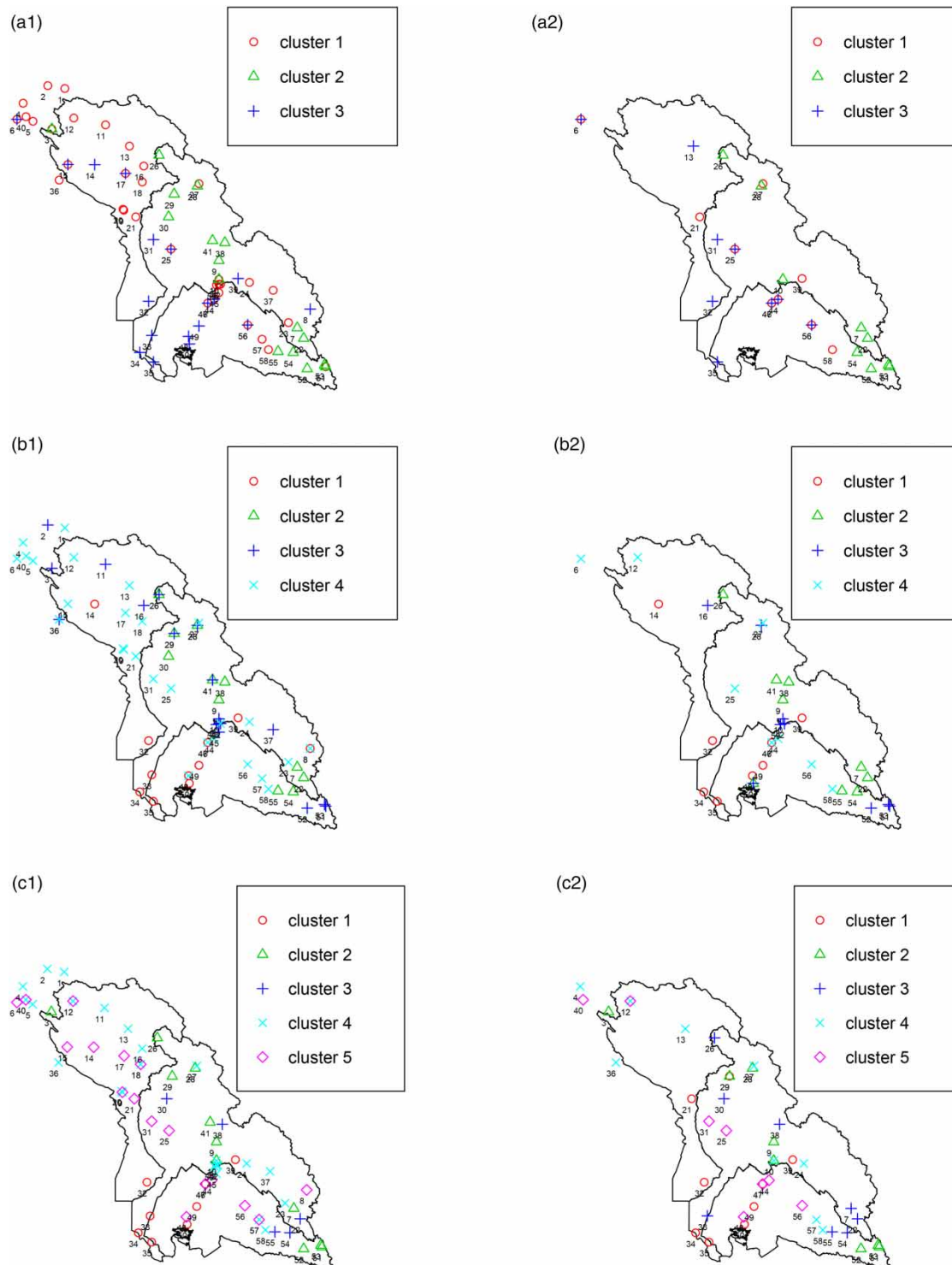
As previously stated, sites in a cluster may not necessarily follow the same frequency distribution. Therefore, it is necessary to apply the discordancy and homogeneity test to each region to assess this issue.



**Figure 2** | RFA steps flowchart.

#### 4.2. Discordancy and homogeneity tests

After the initial clustering, the discordant sites were identified and removed, performing the bivariate discordancy test. After removing the discordant sites, their membership in other clusters was checked. The adjusted clusters,  $\mathbf{a}_2$ ,  $\mathbf{b}_2$ , and  $\mathbf{c}_2$ , shown in Figure 3, were obtained after applying the discordancy test. Adjusting the clusters resulted in a decrease in the density of sites in the regions, particularly in the Karun sub-basin (Figure 3). Moreover, the total number of discordant sites belonging to no region decreased as the number of clusters increased. The results for the adjusted clusters are presented in Table 1. In this study, the results of modified clustering analysis showed that increasing the number of clusters led to a decrease in the total number of removed stations, but did not significantly affect the size of each region. Given the fact that the discordancy test may not necessarily be sufficient to create homogeneous regions, a bivariate homogeneity test was applied to all the regions in all three cases.



**Figure 3** | Dispersion of clusters before (a<sub>1</sub>, b<sub>1</sub>, and c<sub>1</sub>) and after (a<sub>2</sub>, b<sub>2</sub>, and c<sub>2</sub>) adjustment for three different numbers of clusters.

In the case of ( $nc = 3$ ), three homogeneous regions were created, each containing nine sites. In the cases of ( $nc = 4$ ) and ( $nc = 5$ ), two and one heterogeneous regions were detected, respectively. The homogeneity index of each region is presented in Table 1. Since the homogeneity condition is only met when  $nc = 3$ , this study applied regional drought frequency analysis to these three homogeneous regions. The size of each region is related to the concept of degree of homogeneity; smaller



**Table 1** | Adjusted clusters and homogeneity index values for three different clustering modes

	Three clusters			Four clusters				Five clusters				
	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>
	6	7	6	14	7	10	6	32	3	22	4	12
	21	10	25	32	9	16	12	34	9	30	10	25
	25	22	31	34	22	28	25	35	10	38	12	31
	27	26	32	35	26	42	27	39	28	54	13	40
	44	28	35	39	38	52	44	47	51	55	24	44
	46	51	44	47	41	52	46	49	52	33	27	47
	56	52	47	48	54	53	56	50	53	7	36	48
	58	53	56	49	55	59	58	29	29	26	57	56
	39	54	13		50	50		21			58	
H	-1.0	-0.1	0.8	3.8**	0.1	-2.3	-0.8	1.2*	0.8	0.6	0.9	-4**
$\sum n$	293	329	296									

\*possibility homogeneous \*\*definitely heterogeneous

$$Reliability = \frac{\sum n}{5}$$

regions tend to have higher degrees of homogeneity. However, the small size of the region leads to insufficient data for RFA (Chebana & Ouarda 2007). As mentioned before, the aim of RFA is to improve the reliability of estimates by pooling data from multiple sites rather than just one. Larger regions generally have more data, which leads to more reliable estimates. To determine if a region is large enough, the 5T guideline is utilized where T represents the return period (Sadri & Burn 2011).

Notably, the formed regions do not coincide with the national classification of watersheds because each base is different. In other words, regionalization does not necessarily rely on the physical vicinity of weather stations; rather, it is based on the similarity in shape of their probability distributions.

At this stage, observed data from all stations within the region were aggregated and a frequency analysis of data from the entire region was performed.

#### 4.3. Marginal distributions

Five widely used three-parameter probability distribution functions in hydrology, including (1) Generalized Logistic, (2) Generalized Extreme-value, (3) Generalized Normal, (4) Pearson Type III, and (5) Generalized Pareto, were candidates to fit as marginal distribution functions for the two variables of drought severity and duration.

The  $Z^{\text{Dist}}$  statistic was used to calculate the goodness-of-fit of the drought severity. As noted in Table 2, the distribution function statistics for the three-parameter candidates are not within the acceptable range in all three regions. It is practical to use the Wakeby distribution function as the regional distribution for regions that do not completely meet the homogeneity (Hosking & Wallis 1993). Therefore, five-parameter Wakeby distribution was used as a regional distribution. Table 2 also shows the parameters of the distributions.

In terms of duration, the KS statistical method is used in this study. The critical and extracted values are summarized in Table 2 for all three regions. According to that, all candidate drought duration distribution functions are within the acceptable range for all three regions. Therefore, the criterion for choosing the best distribution function is the Root Mean Square Error (RMSE) between theoretical and experimental probabilities. Finally, the Generalized Logistic function was chosen as the best distribution function for all three regions. The parameters of the distribution function chosen for the region are provided in Table 2. These parameters were estimated by L-moments.

#### 4.4. Copula

In this study, the Frank, Clayton, and Gumble families of copulas are candidates for joint probability distribution functions. Based on the maximum pseudo-likelihood results, Clayton was chosen in all three regions. The candidate parameters are also estimated by maximum pseudo-likelihood (Table 3).

**Table 2** | Goodness-of-fit for candidates and the parameters for the best-fitted marginal distribution

Region	Variable		Distribution					Best-fitted distribution	Parameters				
			glo	gev	gno	pe3	gpa						
R <sub>1</sub>	S	Z <sup>Dist</sup>	−3.37	−4.70	−5.00	−5.63	−7.75	wak*	$\xi$	$\alpha$	$\beta$	$\gamma$	$\delta$
									1.785	9.525	3.804	0.919	0.270
		KS	0.28	0.28	0.28	0.28	0.27	glo	$\xi$	$\alpha$	$\kappa$		
R <sub>2</sub>	S	RMSE	0.005	0.009	0.011	0.014	0.024		5.757	1.093	−0.217		
		Z <sup>Dist</sup>	−2.52	−3.71	−4.15	−4.97	−6.61	wak*	$\xi$	$\alpha$	$\beta$	$\gamma$	$\delta$
									1.466	10.935	4.486	1.211	0.242
R <sub>3</sub>	S	KS	0.23	0.23	0.23	0.26	0.31	glo	$\xi$	$\alpha$	$\kappa$		
		RMSE	0.003	0.006	0.008	0.13	0.017		5.697	1.159	−0.279		
		Z <sup>Dist</sup>	−1.65	−3.37	−3.52	−4.00	−7.13	wak*	$\xi$	$\alpha$	$\beta$	$\gamma$	$\delta$
R <sub>3</sub>	D								1.652	9.411	3.61	0.993	0.229
		KS	0.23	0.30	0.30	0.30	0.27	glo	$\xi$	$\alpha$	$\kappa$		
		RMSE	0.004	0.008	0.009	0.011	0.020		5.784	1.056	−0.207		

\*Since no three-parameter distribution was in the acceptable range concerning the goodness-of-fit criterion, the five-parameter Wakeby was fitted to the data as the most appropriate choice

**Table 3** | Maximum pseudo-likelihood and the parameters for the copulas

Region		Frank	Clyton	Gumbel
R <sub>1</sub>	$\theta$	8.938	2.581	2.794
	mlp	0.059	0.139	0.051
R <sub>2</sub>	$\theta$	8.316	2.540	2.658
	mlp	0.059	0.342	0.012
R <sub>3</sub>	$\theta$	8.416	2.434	2.651
	mlp	0.051	0.242	0.025

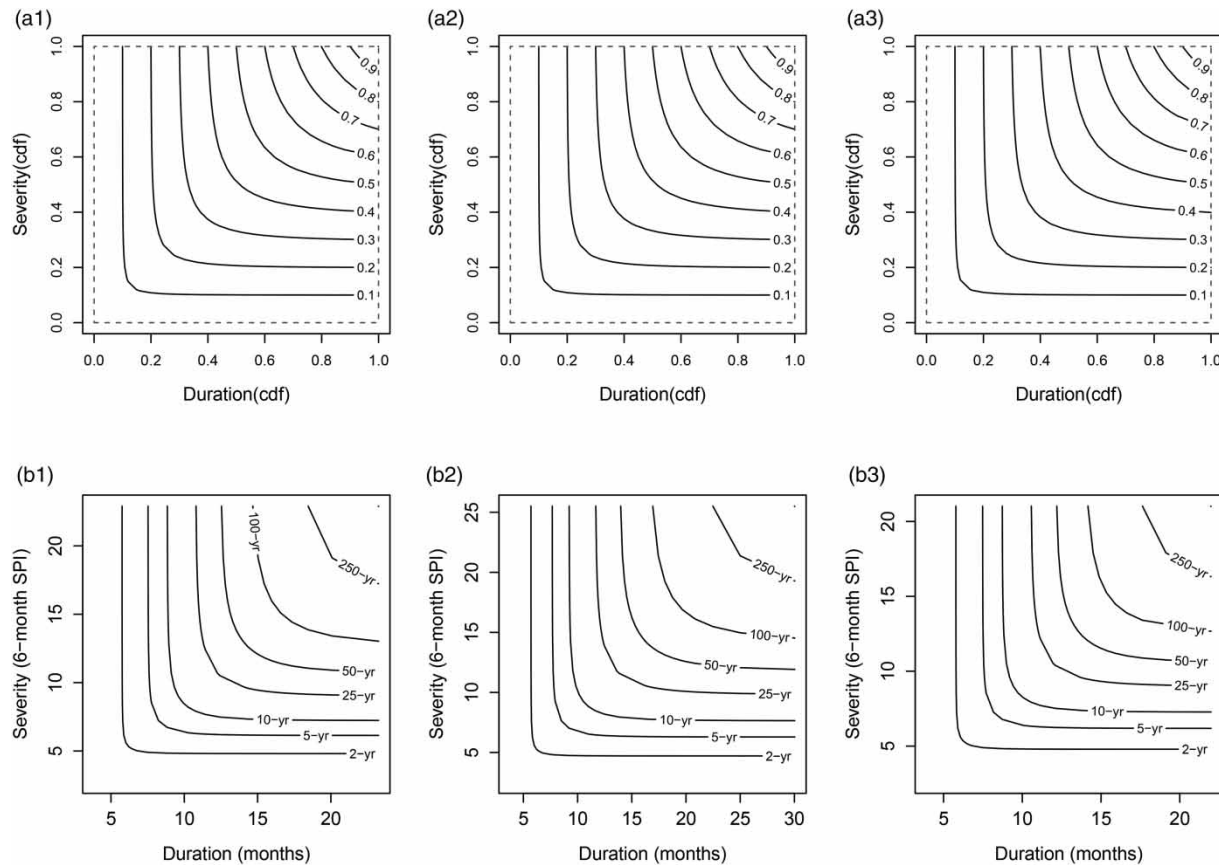
At this stage, the return period for each drought occurrence in the regions was estimated using the marginal distributions and copula. Note that the reliability of the estimated return periods depends on the length of the sample data for each region and is determined by dividing the total length by 5 ( $\frac{n \times N}{5}$ ) (Sadri & Burn 2011). Contours of the joint distribution of drought duration and severity, as well as the bivariate return period for all regions, are illustrated in Figure 4.

The first quantile among the most extreme drought recorded at each site was identified, characterized by a duration of 16 months and a severity index of 12.7. The return period for each site was determined through regional frequency analysis, resulting in values of 63.45, 38.74, and 73.1 for Regions 1, 2, and 3, respectively. This indicates that a drought event with similar duration and severity occurs approximately every 63.5 years in the sites exclusively belonging to Region 1, every 38.7 years in Region 2, and every 73.1 years in Region 3. Notably, the return period of the same drought event in Region 2 is significantly shorter than that in Regions 1 and 3. Also, the results suggest an earlier occurrence in Region 1 compared to Region 3.

However, for the sites belonging to more than one region, such as Site Number 6, which exhibits a membership of 0.6 in Region 1 and 0.35 in Region 2, the return period was computed using a weighted average formula ( $\frac{0.6 \times 63.45 + 0.35 \times 73.1}{0.95} = 67$ ), accounting for the fuzzy clustering.

These estimations are deemed reliable, given that they do not deviate significantly from  $\frac{n \times N}{5}$  for each respective region.

The second (duration = 17, severity = 13.9) and third (duration = 19, severity = 16) quantiles were not pursued, as return periods beyond those presented in Table 4 are deemed unreliable. Notably, the severity of the first quantile for extreme drought events surpasses that of the third quantile for all drought events, with a duration of 7 and a severity of 5.9.

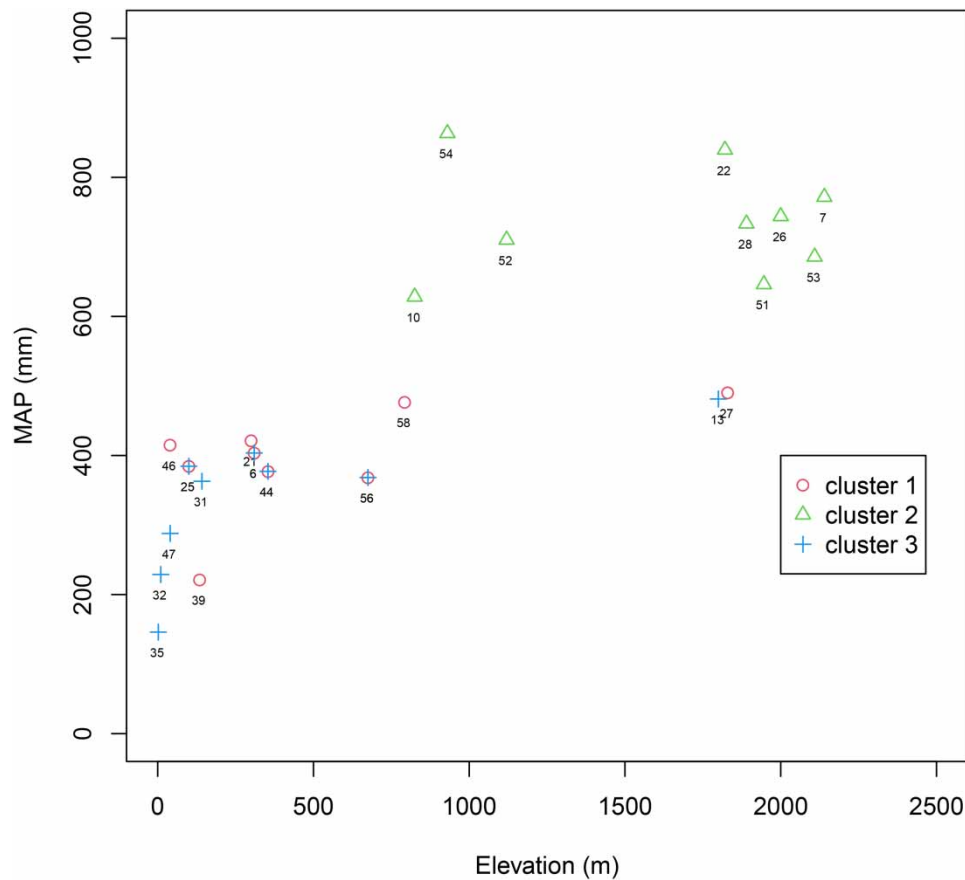


**Figure 4** | Contours of joint probability distribution function for all regions. a1–a3 indicates the probability contours for region 1–3, and identically b1–b3 represent return period contours.

**Table 4** | The regional frequency analysis for all sites. The return period of the median of the most severe drought in all sites, with the duration of 18 months and severity of 13.9, was considered calculated

Site.No.	Membership	Return Period (year)	$\frac{N}{5}$
21-27-58-39-46	Region 1	63.4	63
7-10-22-26-28-51-52-53-54	Region 2	38.7	73
31-32-35-47-13	Region 3	73.1	72
6	0.6 Region 1	67.0	
	0.35 Region 3		
25	0.45 Region 1	68.6	
	0.51 Region 3		
44	0.39 Region 1	69.2	
	0.57 Region 3		
56	0.33 Region 1	69.8	
	0.64 Region 3		

It has been established that the proximity of weather stations does not influence regionalization; instead, regionalization is predicated on the similarity in the shape of their probability distributions. Furthermore, we conducted an examination of the relationship between elevation, MAP, and regionalization. In this study, a correlation of 0.59 was observed between elevation and MAP. Figure 5 illustrates that the sites within Region 2 exhibit noticeably higher elevations and MAPs compared to



**Figure 5** | The elevation and MAP for all three regions.

**Table 5** | Standard deviation of normalized MAP, longitude, and latitude for the three regions

Region	MAP	Longitude	Latitude
R <sub>1</sub>	0.3610587	0.9044799	0.8587007
R <sub>2</sub>	0.3717357	0.7313328	0.9435776
R <sub>3</sub>	0.469	0.784707	0.947184

Regions 1 and 3. Meanwhile, no significant difference in elevation and MAP exists between Regions 1 and 3, which display an overlap and share four common sites. Notably, despite the marginal distribution parameters of these two regions being close, the difference in copula parameters precludes their merging into a single region. Here, it can be asserted that sites with higher MAPs experience a shorter return period.

Assuming that the conclusion is correct, it must be represented in the sites in close proximity while having different MAPs. For instance, Sites 27 and 28 are adjacent in terms of longitude and latitude, whereas the MAP of Site 28, at 733 meters, is higher than that of Site 27, which is 490 meters. Consequently, it is anticipated that Site 28 has a shorter return period. Table 4 illustrates that the return period of Site 28 (38.7) is indeed shorter than that of Site 27 (63.4). Similar findings apply to Sites 54 and 58.

Moreover, an examination of Sites 25 and 31, which are geographically close and have similar MAPs (386 and 362, respectively), reveals return periods of approximately 68.6 and 73.1, respectively, with the former being slightly shorter. These comparisons substantiate the assertion that sites with higher MAPs undergo a shorter return period for the same drought event.

This claim can also be deduced through an analysis of the standard deviation of MAP, Longitude, and Latitude. As depicted in Table 5, the standard deviation of MAP for each group is significantly lower than that of Longitude and Latitude. This observation affirms that sites within a region are more closely associated with similar MAPs than with horizontal distances. It is crucial to emphasize that all inputs underwent normalization and standardization, exhibiting a standard deviation of 1 and an average of 0 before the clustering analysis.

## 5. LIMITATIONS

In the course of this investigation, a substantial number of sites were identified that did not belong to any region. Consequently, there arises an imperative need to explore innovative strategies aimed at minimizing the exclusion of such sites.

It is crucial to emphasize that the reliability of this approach cannot be corroborated by at-site frequency analysis results, primarily due to the limited length of the at-site records. The short length of these records precludes conclusive confirmation of the obtained results.

To fortify the robustness of our model, alternative validation methods, including data-driven models and machine learning approaches, warrant exploration. Furthermore, expanding the application of this methodology to larger geographical scales, such as continental or global domains, would provide valuable insights. These propositions present intriguing avenues for future investigations, and their pursuit would contribute significantly to the refinement and generalizability of the findings.

## 6. CONCLUSION

This study utilizes bivariate RFA to estimate the return period of a severe drought in Iran across three catchments (Karkheh, Karun, and Jarahi-Zohreh) with increased reliability. The fuzzy c-means clustering method was used to cluster 59 meteorological stations into regions, followed by bivariate discordancy and homogeneity tests using multivariate L-moments to investigate the homogeneity of the regions.

It was observed that the total number of discordant sites was higher for  $nc = 3$  than for  $nc = 4$ . It was similarly higher for  $nc = 4$  compared to  $nc = 5$ . Thus, as expected, increasing the number of clusters reduces the total number of eliminated sites. In fact, when fewer clusters are considered, there would be more sites that do not belong to any of the clusters, but it does not necessarily lead to more or fewer sites in each cluster. Although it may be expected that as the number of clusters increases, the size of regions decreases, it did not vary significantly for any clusters. Therefore, changing the number of clusters does not necessarily affect region size.

Only for the  $nc = 3$  clustering, all regions were homogeneous. The results showed that combining the initial clustering method and discordancy test does not necessarily lead to the formation of homogeneous regions, which is a serious weakness. In this study, the three-parameter Generalized Logistic and the five-parameter Wakeby were the best-fitted distribution functions for the regional drought duration and severity variables, respectively. Regarding copula and drought duration, the regions were homogeneous. However, it was not the case regarding the severity since no three-parameter marginal distribution function was fitted to the drought severity variable in any of the regions. In other words, satisfying the bivariate homogeneity index does not necessarily result in region homogeneity regarding the marginal distribution functions. However, the homogeneity of the marginal distribution function is still necessary. Clayton was the most appropriate copula function for all three regions.

The approach of regional frequency analysis, focusing on probability distribution similarities rather than physical proximity, has provided reliable estimations for each region. The distinct temporal patterns in return periods across Regions 1, 2, and 3 underline the variability of drought patterns.

Given the lower return period in Region 2 compared to Regions 1 and 3, it is deduced that sites with higher MAP and, correspondingly, higher elevation are more prone to experiencing shorter return periods of same drought events, in contrast to sites with lower MAP or lower elevation.

In summary, this study advances our understanding of regionalized drought patterns. Still, further investigations with extended data and alternative methodologies are necessary for a more comprehensive grasp. The findings hold practical significance for policymakers and researchers alike, underscoring the importance of considering regional nuances in drought risk assessments and water resource management strategies.



However, it's crucial to acknowledge limitations, particularly the reliance on regional frequency analysis due to the short duration of at-site records. This underscores the need for future research with extended datasets to enhance the robustness of our conclusions.

## CODE AVAILABILITY

The code was written in R programming by the original author for the Masters thesis at Shahid Beheshti University in Iran. The availability of the code requires the consent of the other authors and the university.

## DATA AVAILABILITY STATEMENT

All relevant data are available in the supplementary material.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

- Abbasian, M. S., Najafi, M. R. & Abrishamchi, A. 2021 Increasing risk of meteorological drought in the Lake Urmia basin under climate change: introducing the precipitation–temperature deciles index. *Journal of Hydrology* **592**, 125586.
- Abdi, A., Hassanzadeh, Y., Talatahari, S., Fakheri-Fard, A. & Mirabbasi, R. 2017 Regional bivariate modeling of droughts using l-comoments and copulas. *Stochastic Environmental Research and Risk Assessment* **31** (5), 1199–1210.
- Ahani, A., Nadoushani, S. S. M. & Moridi, A. 2020 Regionalization of watersheds by finite mixture models. *Journal of Hydrology* **583**, 124620.
- Ahani, A., Nadoushani, S. S. M. & Moridi, A. 2022 A ranking method for regionalization of watersheds. *Journal of Hydrology* **609**, 127740.
- Amoli, A. A., Naeini, A. S. & Ravan, S. H. 2015 Effective Use of Space-Based Information to Monitor Disasters and Its Impacts. Lessons Learnt from Drought in Iran. Un-Spider, Tehran. [https://www.un-spider.org/sites/default/files/Iran\\_booklet\\_0905\\_final\\_small.pdf](https://www.un-spider.org/sites/default/files/Iran_booklet_0905_final_small.pdf).
- Azam, M., Park, H. K., Maeng, S. J. & Kim, H. S. 2018 Regionalization of drought across South Korea using multivariate methods. *Water* **10** (1), 24.
- Bezdek, J. C., Ehrlich, R. & Full, W. 1984 FCM: the fuzzy c-means clustering algorithm. *Computers & Geosciences* **10** (2-3), 191–203.
- Chebana, F. & Ouarda, T. B. M. J. 2007 Multivariate L-moment homogeneity test. *Water Resources Research* **43** (8), 1–14.
- Chebana, F., Ouarda, T. B. M. J., Bruneau, P., Barbet, M., El Adlouni, S. & Latraverse, M. 2009 Multivariate homogeneity testing in a northern case study in the province of Quebec, Canada. *Hydrol Process* **23** (2), 1690–1700.
- Dunn, J. C. 1974 Some recent investigations of a new fuzzy partitioning algorithm and its application to pattern classification problems. *Journal of Cybernetics* **4** (2), 1–15.
- Gocic, M. & Trajkovic, S. 2013 Analysis of precipitation and drought data in Serbia over the period 1980–2010. *Journal of Hydrology* **494**, 32–42.
- Goyal, M. K. & Sharma, A. 2016 A fuzzy c-means approach regionalization for analysis of meteorological drought homogeneous regions in Western India. *Natural Hazards* **84** (3), 1831–1847.
- Hosking, J. R. M. & Wallis, J. R. 1993 Some statistics useful in regional frequency analysis. *Water Resources Research* **29** (2), 271–281.
- Hosking, J. R. M. & Wallis, J. R. 1997 *Regional Frequency Analysis: an Approach Based on L-Moments*. Cambridge University Press, Cambridge.
- Karahacane, H., Meddi, M., Chebana, F. & Saaed, H. A. 2020 Complete multivariate flood frequency analysis, applied to northern Algeria. *Journal of Flood Risk Management* **13** (4), e12619.
- Masselot, P., Chebana, F. & Ouarda, T. B. M. J. 2017 Fast and direct nonparametric procedures in the L-moment homogeneity test. *Stochastic Environmental Research and Risk Assessment* **31** (2), 509–522.
- McKee, T. B., Doesken, N. J. & Kleist, J. 1993 The relationship of drought frequency and duration to time scales. In: *Proceedings of the 8th Conference on Applied Climatology* Vol 17. Boston, pp. 179–183.
- Mirabbasi, R., Fakheri-Fard, A. & Dinpashoh, Y. 2012 Bivariate drought frequency analysis using the copula method. *Theoretical and Applied Climatology* **108** (1-2), 191–206.
- Montaseri, M., Amirataee, B. & Rezaie, H. 2018 New approach in bivariate drought duration and severity analysis. *Journal of Hydrology* **559**, 166–181.
- Mortuza, M. R., Moges, E., Demissie, Y. & Li, H.-Y. 2019 Historical and future drought in Bangladesh using copula-based bivariate regional frequency analysis. *Theoretical and Applied Climatology* **135**, 855–871.
- Nabaei, S., Sharafati, A., Yaseen, Z. M. & Shahid, S. 2019 Copula based assessment of meteorological drought characteristics: regional investigation of Iran. *Agricultural and Forest Meteorology* **276**, 107611.
- Nelsen, R. B. 2007 *An Introduction to Copulas*. Springer Science & Business Media, New York.
- Ngongondo, C. S., Xu, C.-Y., Tallaksen, L. M., Alemaw, B. & Chirwa, T. 2011 Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches. *Stochastic Environmental Research and Risk Assessment* **25** (7), 939–955.

- Parvizi, S., Eslamian, S., Gheysari, M., Gohari, A. & Kopai, S. S. 2022 Regional frequency analysis of drought severity and duration in Karkheh River Basin, Iran using univariate L-moments method. *Environmental Monitoring and Assessment* **194** (5), 336.
- Portela, M. M., dos Santos, J. F., Silva, A. T., Benitez, J. B., Frank, C. & Reichert, J. M. 2015 Drought analysis in southern Paraguay, Brazil and Northern Argentina: regionalization, occurrence rate and rainfall thresholds. *Hydrology Research* **46** (5), 792–810.
- Rahman, A. S. & Rahman, A. 2020 Application of principal component analysis and cluster analysis in regional flood frequency analysis: a case study in New South Wales, Australia. *Water* **12** (3), 781.
- Rajsekhar, D., Mishra, A. K. & Singh, V. P. 2013 Regionalization of drought characteristics using an entropy approach. *Journal of Hydrologic Engineering* **18** (7), 870–887.
- Raziei, T., Bordi, I. & Pereira, L. S. 2008 A precipitation-based regionalization for Western Iran and regional drought variability. *Hydrology and Earth System Sciences* **12**, 1309–1321.
- Raziei, T., Mofidi, A., Santos, J. A. & Bordi, I. 2012 Spatial patterns and regimes of daily precipitation in Iran in relation to large-scale atmospheric circulation. *International Journal of Climatology* **32** (8), 1226–1237.
- Sadri, S. & Burn, D. H. 2011 A fuzzy c-means approach for regionalization using a bivariate homogeneity and discordancy approach. *Journal of Hydrology* **401** (3–4), 231–239.
- Sadri, S. & Burn, D. H. 2014 Copula-based pooled frequency analysis of droughts in the Canadian Prairies. *Journal of Hydrologic Engineering* **19** (2), 277–289.
- Serfling, R. & Xiao, P. 2007 A contribution to multivariate L-moments: L-comoment matrices. *Journal of Multivariate Analysis* **98** (9), 1765–1781.
- Shamshirband, S., Gocić, M., Petković, D., Javidnia, H., Ab Hamid, S. H., Mansor, Z. & Qasem, S. N. 2015 Clustering project management for drought regions determination: a case study in Serbia. *Agricultural and Forest Meteorology* **200**, 57–65.
- Shiau, J. T. 2006 Fitting drought duration and severity with two-dimensional copulas. *Water Resources Management* **20** (5), 795–815.
- Shiau, J.-T. & Modarres, R. 2009 Copula-based drought severity-duration-frequency analysis in Iran. *Meteorological Applications: A Journal of Forecasting, Practical Applications, Training Techniques and Modelling* **16** (4), 481–489.
- Shiau, J.-T. & Shen, H. W. 2001 Recurrence analysis of hydrologic droughts of differing severity. *Journal of Water Resources Planning and Management* **127** (1), 30–40.
- Sklar, A. 1973 Random variables, joint distribution functions, and copulas. *Kybernetika* **9** (6), 449–460.
- Svoboda, M., Hayes, M. & Wood, D. 2012 Standardized precipitation index user guide. *World Meteorological Organization Geneva, Switzerland*.
- Xu, K., Yang, D., Yang, H., Li, Z., Qin, Y. & Shen, Y. 2015 Spatio-temporal variation of drought in China during 1961–2012: a climatic perspective. *Journal of Hydrology* **526**, 253–264.
- Yevjevich, V. M. 1967 Objective Approach to Definitions and Investigations of Continental Hydrologic Droughts, an. Hydrology papers (Colorado State University); no. 23.
- Yoo, J., Kwon, H.-H., Kim, T.-W. & Ahn, J.-H. 2012 Drought frequency analysis using cluster analysis and bivariate probability distribution. *Journal of Hydrology* **420**, 102–111.
- Yu, K.-x., Xiong, L. & Gottschalk, L. 2014 Derivation of low flow distribution functions using copulas. *Journal of Hydrology* **508**, 273–288.
- Zhang, Q., Qi, T., Singh, V. P., Chen, Y. D. & Xiao, M. 2015 Regional frequency analysis of droughts in China: a multivariate perspective. *Water Resources Management* **29** (6), 1767–1787.

First received 6 September 2023; accepted in revised form 29 December 2023. Available online 1 February 2024