Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

Regional Flood Frequency Analysis of Northern Iran

Kuzey İran'ın Bölgesel Taşkın Frekans Analizi

Maryam Adhami

Department of Watershed Management Engineering, Faculty of Natural Resources, Tarbiat Modares University, Noor, Mazandaran, Iran
*Corresponding Author / Sorumlu Yazar *:* m.adhami66@yahoo.com

Abstract

The combination of the L-moment approach and multiple regression offers an attractive solution to provide flood estimation at ungauged sites within the Gorganrood and Ghare-sou river basins in the north of Iran. This research has two main goals including regionalization by cluster analysis and regional estimation of flood quantile at the site of interest. After data analysis regarding climatic and hydrologic data series, hierarchical approach was carried out to identify homogeneous regions. The homogeneity test was done by H-Statistic, a testing method based on L-moments. The results showed that a subdivision of selected watersheds into homogenous groups is necessary; therefore, two homogenous regions were formed. In the present study, five three-parameter distributions were fitted to the homogeneous regions and the best-fit one was identified using the L-moments approach. The results of the goodness-of-fit analysis for the two regions introduced the Generalized Pareto (GPA) distribution for both regions as acceptably close fits to the regional average L-moments. Besides, multiple regression was applied to diagnose the effective independent parameters on discharge value. The results reported percent of permeable formations, average annual precipitation, and stream slope as the most effective variables.

Keywords*: Clustering, Gorganrood and Ghare-sou, L-moments, Principal component analysis, Regional flood frequency analyses, Regionalization*

Öz

L-moment yaklaşımı ve çoklu regresyonun kombinasyonu, İran'ın kuzeyindeki Gorganrood ve Ghare-sou nehri havzalarındaki ölçüm olmayan alanlarda taşkını tahmin etmek için cazip bir çözüm sunmaktadır. Bu araştırma, kümeleme analizi ile çalışma alanının bölgeselleştirilmesi ve taşkın kuantillerinin bölgesel tahmini olmak üzere iki ana amaca yöneliktir. Verilerin analizinden sonra homojen bölgeleri belirlemek için hiyerarşik bir yaklaşım gerçekleştirilmiştir. Homojenlik testi, L-momentlerine dayalı bir test yöntemi olan H-Statistic ile yapılmıştır. Sonuçlar, seçilen havzaların homojen gruplara bölünmesinin gerekli olduğunu göstermiştir; dolayısıyla iki homojen bölge oluşmuştur. Bu çalışmada, homojen bölgelerde beş adet üç parametreli dağılımın uyumluluğu incelenmiş ve en uygun dağılım L-momentler yaklaşımı kullanılarak belirlenmiştir. İki bölge için uyumluluk analizinin sonuçları, her iki bölge için de Generalized Pareto (GPA) dağılımını, bölgesel ortalama L-momentlerine kabul edilebilir ölçüde yakın olduğunu göstermiştir. Ayrıca debi üzerindeki etkili bağımsız değişkenlerin tespiti için çoklu regresyon uygulanmıştır. Sonuçlar geçirgen formasyonların yüzdesi, yıllık ortalama sıcaklık ve akarsu eğiminin en etkili değişkenler olduğunu göstermektedir.

Anahtar Kelimeler: *Kümelendirme, Gorganrood ve Ghare-suo, L-moments, Temel bileşenler Analizi, Bölgesel taşkın frekansı analizi, Bölgeselleştirme*

1. Introduction

Flood frequency analysis performs the estimation of the return period and flood magnitude [1]. This process is often essential for the design of hydraulic structures such as dams and bridges, and for hydrological applications such as dam safety analyses and reservoir management. This information is required at watersheds where stream flow measurements are not long enough to provide a precise calculation of the flood magnitude, frequency distribution, and the long return periods estimation of flood, or where there is no data at all [2]. In arid and semi-arid regions flood frequency analysis is encountering data and information shortage issues. Especially, in these regions, the reliable estimate is not possible due to short records [3]. Arid and semi-arid climatic condition covers more than 75% of Iran where despite the low annual precipitation, large floods occur sometimes. Using records from a similar region regarding flood behavior is an important and practical way to provide more information, rather than only at-site data [4].

On the other hand, geomorphologic and hydrologic parameters with characteristics of the watersheds are necessary to enable better planning and conducting the proper strategies for management practices [5]. Developed strategies could be implemented in a prioritized manner, due to the involvement of huge investment. Toward this aim, it is expected to rehabilitate the watershed (s) systematically and sustainably [6]. Scholars in related fields have claimed that regionalization of watersheds is essential to develop regional flood flow equations which would be used to estimate the magnitudes of flood at locations that suffer from actual flood data shortage. Regional flood frequency analysis provides a solution to such a problem and has widely been used [4, 7-11]. The analysis uses spatial data to compensate for the lack of temporal data, accomplished in a region with similar flood behavior. The background assumption is that flood data regarding the homogeneous region without considering a scaling factor are calculated from the same frequency distribution. The method involves two major steps; the first one is the identification of homogeneous watersheds from the

DOI:10.21205/deufmd.2024267711

Atıf şekli / How to cite:

Adhami, M. (2024). Regional flood frequency analysis of northern Iran. DEUFMD, 26(77), 272-280.

Geliş Tarihi / Received: 09.08.2023
Kabul Tarihi / Accepted: 04.10.2023

hydrological perspective, which is performable by selecting gauging stations that constitute a region with sufficient homogeneity of flow characteristics. The second step forms calculating regional equations for each group [12]. The homogeneous regions were simply defined according to the geographic proximity of the different watersheds. A significant necessity for regional flood frequency analysis is the delineation of the region used for the flow information transfer. A region, in this perspective, is considered a collection of watersheds that are similar in terms of watersheds hydrologic behavior, not necessarily geographically adjacent. This step is the most challenging but important step in regional flood frequency analysis. Various regionalization techniques have been developed by researchers for the determination of homogeneous regions [7, 13-21]. Hierarchical Agglomerative Clustering has become a popular tool for regional distribution identification, and testing of outlier stations [22-27]. Tasker and Thomas (1978) applied geographical regionalization method in Arkansas [22]. Regional equations regarding 4 regions have been presented. Acreman and Sinclair (1986) used Ward's method for the selection of parent distribution to fit extreme streamflow data of some sites in Scotland [23]. Between five homogeneous regions just one group could not show the discharge variations. Stamy and Hess (1993) carried out a regional flow frequency analysis in Florida and Georgia in America using Ward's method [24]. They classified the study area in four groups. All four groups have presented significant results. Kjeldsen and Smithers (2002) also investigated the spatial variation and regional frequency distribution for KwaZulu-Natal provinces, South Africa, using the index-flood method. According to the results, two clusters were distinguished [25]. Farsadnia et al. (2014) applied a two-level Self-Organization Feature Map (SOM) and three clustering methods (fuzzy c-mean, K-mean, and Ward's Agglomerative hierarchical clustering) to identify hydrologically homogeneous regions in Mazandaran province of Iran [26]. Unified distance matrix algorithm and mentioned clustering methods formed regions for flood frequency analysis. Four regions were achieved using a heterogeneity test. The results suggested that the combination of SOM and Ward is much better than the combination of either SOM or fuzzy c-mean and K-mean. Calegario et al. (2020) conducted regionalization to identify Hydrologically Homogeneous Regions (HHRs) in the Doce River basin [27]. Grouping was done based on geographical convenience methods and cluster analysis. Six statistical indices were used to assess regionalized flows. It was concluded that such a physical analysis reduced the subjectivity in the identification of HHRs.

The essence of the regionalization process is to define similar watersheds to ensure extreme flow information transfer of sites within the region. The main objectives of this paper are summarized in three steps, the first one is to determine the homogeneous regions, the second one is to estimate regionalized parameters, and the last one is calculating the best equations for each region in different period return time in Gorganrood and Ghare-sou watersheds, Golestan province of Iran.

2. Material and Method

This phase involves four main stages: (i). screening the data and determining the main site and at-site characteristics that affect the flood magnitude and applying them in the regionalization of the study area, (ii). identifying homogeneous regions by cluster analysis and region of-influence methods, testing the homogeneity of regions, (iii). investigating the best-fit distribution for the study area based on L-moments approaches and (iv). calculating estimation equations for each region by multiple regression.

2.1. Study area

The Gorganrood and Ghare-sou Rivers originate from the Gorgan plate and flow into the Khazar Sea. The rivers have a drainage area of 11786 km^2, and a mean annual temperature of 17° C. Most of the basin's area is humid or semi-humid with a mean annual precipitation of 750 mm. The length of the Ghare-Sou and Gorganrood rivers are 160 and 300 km, respectively. All streams flowing through study watersheds have a constant regime. The mean annual discharge of the streams ranges among 0.32-12.54 m^3/s. Figure 1 represents the location of the study area and flow gauging stations.

2.2. Selection of the stations

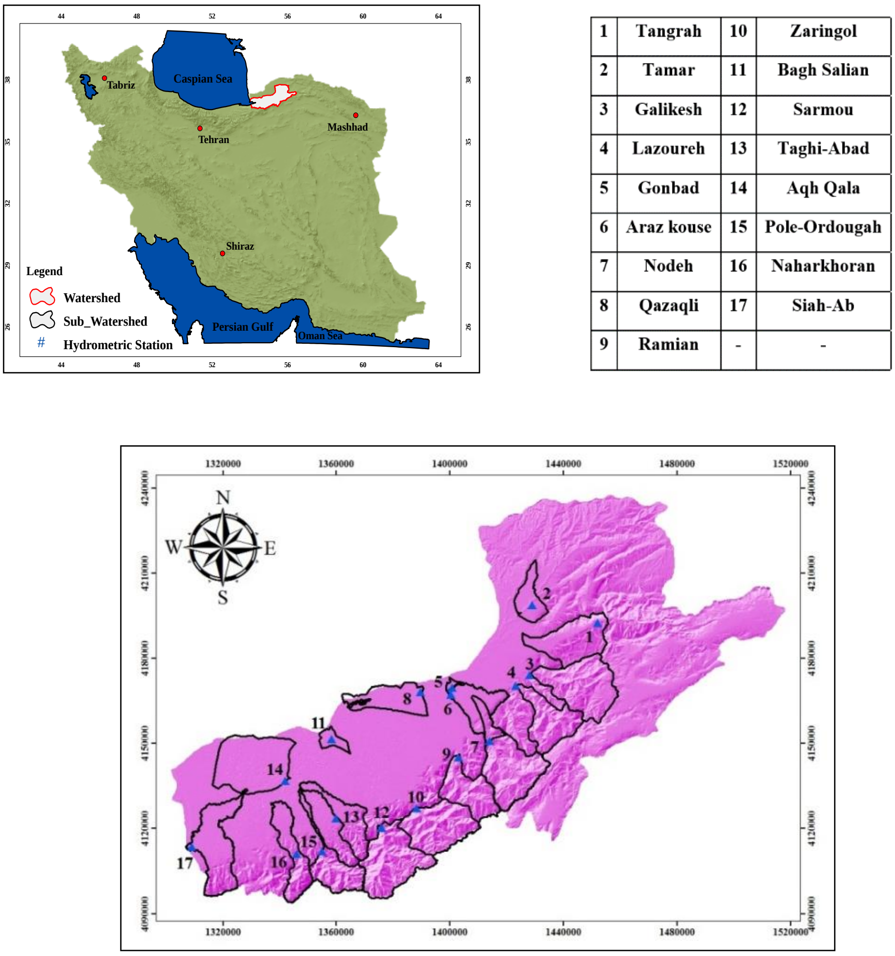
Some important criteria named record length, number of missing data, location of the station, watershed size, and regulation level were considered for station selection. Data series were screened to ensure they include no more missing data. Stations on sub-watersheds with major dams or reservoirs were excluded since discharge magnitude can be greatly affected by regulation in rivers [28-29]. All 17 selected hydrometric stations have 24 years or more of discharge data and are appropriate for study. Daily discharge series measured at the gauging stations were used for this study. In cases with abnormal data, they were supposed outliers, and Grubbs-Beck method was applied to verify them (30). Besides, regression method was used to fill in missing data in the time series of selected watersheds. Some statistical features related to study sub-watersheds discharge values are summarized in Table 1.

Table 1. Statistical features of study station's discharge

Sub-watershed	Minimum	Maximum	Mean	Standard Deviation	Skewness coefficient
Tangrah	0.31	10.39	1.94	1.93	3.88
Tamar	0.81	3.91	1.67	0.66	2.08
Galikesh	1.31	3.95	2.48	0.75	0.38
Gonbad	2.58	14.48	7.36	3.17	0.71
Lazoureh	1.01	3.96	2.09	0.69	0.90
Araz kouse	2.26	9.08	5.46	1.60	0.44
Bagh Salian	0.23	6.90	3.00	1.79	0.61
Taghi-Abad	0.09	0.85	0.43	0.19	0.30
Agh Qala	0.68	23.03	11.73	6.62	0.27
Naharkhoran	0.13	0.71	0.32	0.15	1.00
Siah-Ab	0.21	4.75	2.03	1.01	0.58
Pole-Ordougah	0.17	0.90	0.49	0.20	0.22
Sarmou	0.36	2.48	1.11	0.58	0.94

2.3. Independent parameters

A Geographical Information System (GIS) database was created to retrieve land cover, topography, soil types, and land use, deposition of sediments, and water resources. Watershed attributes were chosen based on their availability and the previous studies on national projects (31-32). These attributes could facilitate the deterioration of the watershed, if do not have a suitable level [30]. Several layers of spatial information about elevation, land cover, soil type, and climate were integrated to retrieve watershed characteristics. Some small watersheds that were not included in these reference layers were manually delineated using a Digital Elevation Model (DEM) at a small scale. DEM at a resolution of 5 meters was used for elevation and slope data. Climate data were obtained from synoptic stations' measured data.



1	Tangrah	10	Zaringol
2	Tamar	11	Bagh Salian
3	Galikesh	12	Sarmou
4	Lazoureh	13	Taghi-Abad
5	Gonbad	14	Aqh Qala
6	Araz kouse	15	Pole-Ordougah
7	Nodeh	16	Naharkhoran
8	Qazaqli	17	Siah-Ab
9	Ramian	-	-

Figure1. The geographical location of Gorganrood and Ghare-sou watersheds Golestan province, Iran

Synoptic stations with 30 years (or more) of data were selected. Several climate parameters such as De-Martonne coefficient, mean annual precipitation and mean annual temperature were computed. All casework parameters (taken from the Natural Resource Administration of Gorgan) are arranged in Table 2.

2.4. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate technique that contains approaches that consider all the variables at the same time. Such approaches focus on the relationships among variables with the individual characteristics of each one [33-34]. PCA was used to treat stream discharge data from 17 stations analyzed for 30 variables in the Gorganrood and Ghare-sou watersheds, North of Iran. The results have identified the effective parameters on stream flow peaks and the variables that were not contributing to these streams' discharge.

In the first step, the covariance matrix is calculated. If X in a matrix format is the original dataset, which embraces m rows (different measurements of a specific attribute) and n columns (which represent the attributes), the covariance matrix C_X will be:

$$C_X = \frac{1}{n} X X^T \tag{1}$$

Where X^T is a transpose matrix of X. Afterward, eigenvectors and eigenvalues are computed. The eigenvector $\vec{v}$ is defined as:

$$C_X \vec{v} = \lambda \vec{v} \tag{2}$$

Where λ is a scalar value, i.e., the eigenvalue. The below equations clarify the steps to solve for eigenvalue and eigenvector:

$$C_X \vec{v} - \lambda \vec{v} = 0 \tag{3}$$

$$\vec{V}(C_X - \lambda I) = 0 \tag{4}$$

$$\mathrm{Det}(C_X - \lambda I) = 0 \tag{5}$$

Where I is the identity matrix of the same dimension as C_X. Consequently, each eigenvector is produced by each λ times $\vec{v}$ which is called the principal component. The dimension of the dataset is equal to the number of principal components. The number of PCs is usually based on the number of eigenvalues greater than 1. The ratio between the sum of the eigenvalues and the eigenvalue of a component shows the percent of the variance in the original dataset represented by that component [35].

2.5. Regionalization method

Cluster Analysis (CA) assembles objects based on their characteristics and contains a group of multivariate techniques whose primary purpose is to find out objects' common features [36]. Cluster analysis classifies objects based on similarity among them based on a predetermined selection criterion. The final groups should contain objects with high internal homogeneity and high external heterogeneity. Hierarchical agglomerative clustering is known as the most popular method and acts in a manner to find intuitive similarity relations among a single sample and whole data set, and provides a dendrogram (tree diagram) [37]. Cluster analysis performs in two ways: model-oriented and distance-oriented [38]. Presently, distance-oriented methods are preferred because unlike the model-based methods, these methods do not take into account the statistical distribution of the data and also do not estimate the parameters of the statistical distribution along with the hidden variable (which is used as the label of the clusters in the introduced model). They are also easy to use and precious in clustering. These methods themselves are classified into two groups: chance models and ordinal ones. Ordinal models are frequently used in comparison with chance models. Ordinal models start with groups whose number is equal to the number of parameters. In other words, there is only one parameter in every single group. In the next step, the more similar groups join together. In the following, these groups join other similar groups. Finally, only one group exists that includes all parameters [39]. In the present study Ward's method as one of the hierarchical clustering approaches is used.

Table 2. Independent parameters of Gorganrood and Ghare-sou watersheds in Golestan province, Iran

Sub-watersheds	Taghi-Abad	Galikesh	Lazoureh	Pole-OrdOugah	Gonbad	Tangrah	Bagh Salian	Naharkhoran	Agh Qala	Siah-Ab	Tamar	Sarmou	Araz kouse
Min-e (m)	-54	-80	200	0	17	100	38	0	-93	-52	80	500	17
Max-e (m)	2538	70	2500	3100	52	1800	-94	3000	85	2612	670	3600	2878
L (km)	34.38	35.28	26.73	41.76	4.89	29.87	12.72	35.91	32.56	39.59	21.88	29.69	47.28
P (km)	127.14	147.16	86.26	145.21	14.97	87.96	34.39	129.25	87.29	164.81	58.24	87.96	205.53
A (km^2)	194.13	264.81	260.80	215.15	95.35	305.55	119.6	212.88	458.93	345.76	122.75	302.40	78.1
G-co	0.092	0.078	0.047	0.09	0.2	0.04	0.1	0.085	0.026	0.066	0.066	0.12	0.05
S-L (km)	34.98	36.21	25.77	35.15	5.11	34.22	12.82	44.04	37.15	43.5	17.59	28.4	78.5
S-S (%)	6.22	0.12	5.9	7.93	0.14	1.33	0.19	7.66	0.05	4.2	1.00	10.92	4.84
T_C (h)	2.98	13.96	2.41	6.32	2.91	5.32	5.25	3.29	19.98	4.11	3.54	2.05	6.13
V (m/s)	3.26	0.72	2.97	0.66	0.49	1.79	0.68	3.72	0.52	2.94	1.38	3.85	3.56
B-S (%)	21.38	3.31	37.6	55.85	9.04	9.09	3.82	13.46	1.37	4.42	5.71	18.95	30.65
P-A (km^2)	52.31	151.78	72.72	72.46	6.15	128.66	14.52	66.71	453.38	133.58	47.17	7.24	35.75
N-A (km^2)	56	37.15	62.2	0	1.24	73.02	11.51	55.69	1.2	67.78	29.21	31	6.29
S-A (km^2)	13.47	28.91	49.56	31.8	0.88	17.95	9.65	55.72	0.01	57.62	15.59	6.49	13.59
E-A (km^2)	7.11	32.21	39	53.34	1.01	57.05	8.77	21.01	0.001	39.86	22.42	45.89	18.93
W-A (km^2)	65.91	14.42	31.17	55.74	0.25	16.16	3.4	10.39	5.51	133.58	8.36	17.4	36.33
D-d1 (km/km^2)	0.34	0.27	0.23	0.26	0.19	0.3	0	0.32	0.29	0.4	0.62	0.56	0.12
D-d2 (km/km^2)	0.26	0.16	0	0.11	0.69	0.14	0	0.14	0.15	0.15	0.32	0.32	0
D-d3 (km/km^2)	0.06	0.05	0	0.06	0.23	0.03	0	0.17	0.08	0.09	0.19	0.14	0
D-d4 (km/km^2)	0.01	0	0	0	0.05	0.05	0	0	0	0.03	0.05	0.11	0
TD_d (km/km^2)	0.66	0.49	0.23	0.42	1.17	0.52	0.07	0.64	0.52	0.66	1.17	1.13	0.12
F (km)	1.69	166.82	4.16	0.42	1.2	1.07	1.33	25	12.66	7.14	7.56	10	6.38
Ra (km)	80.39	0.5	109.32	46.68	8.34	115.89	42.57	22.1	332.62	244.87	107.66	42	189.2
Ag (km)	0	97.49	0.2	0	0	0	3.95	14	113.66	0.16	4.79	0.15	132.43
OT (km)	112.05	0	142.26	168.05	0	188.6	0	150	0	93.53	2.74	50	249.69
P-F (%)	4.34	5.47	1.82	42.65	3.48	0.29	10.94	13.57	53.99	2.35	8.92	0.25	3.93
C-F (%)	4.37	27.08	4.76	2.24	33.9	1.44	7.14	8.75	10.85	0.21	29.44	36.92	17.98
D-C	8.62	7.13	5.55	6.04	7	6	8.21	8.26	9.2	9.68	6	5.4	6.00
A-T (˚c)	16	17	13	13	17	15	17	15	17	17	17	9	13
A-P (mm)	580.11	815.08	873.35	370.36	456.82	772.54	379.04	743.56	418.94	539.34	582.42	778.93	449.99

Min-e: minimum elevation, Max-e: maximum elevation, L: length, P: perimeter, A: area, G-co: Gravelius coefficient, S-L: stream's length, S-S: stream slope, Tc: concentration time, V: velocity of flow, B-S: basin's slope, P-A: area of the pediment, N-A: north land's area, S-A: south land's area, E-A: east land's area, W-A: west land's area, D-d1,2,3,4:1,2,3 and 4 ranked stream's density. TD-d: total density drainage, F: forest's area, Ra: ranches' area, Ag: agriculture field's area, OT: other land uses area, P-F: Percentage of permeable formations, C-F: carbonic bed rocks (namely percentage of impermeable formations), D-C: Demarten coefficient, A-T: annual average temperature, A-P: average annual precipitation.

Based on the rules of mentioned method, grouping is done based on intra-group minimum and inter-group maximum variance [40-41]. Ward's algorithm [42] is one of the common regionalization techniques in climatology and hydrology fields [16, 23, 43-47]. This method acts based on the assumption that information loss or alteration of the value of the objective function that occurs in cluster merging depends only on the relationship between the two merged clusters and is independent of the relationships with any other clusters.

2.6. Cluster validation indices

There are several methods to determine the heterogeneity in cluster analysis; one of them is L-moment. Hosking (1990) introduced L-moments, which are linear combinations of Probability Weighted Moments (PWMs) and can be directly interpreted as the shape of probability distributions and scale measures [48]. Vogel and Fennessey (1993) clarified the advantages of L-moments compared with conventional moments [49]. L-moments and L-moment ratios are the basic concepts for all L-moments steps such as; the homogeneity test, discrimination of heterogeneous regions (discordancy test), best distribution determination for each cluster, and the estimation of parameters (location (n), scale (r), and shape (k)) for selected distributions [4]. According to Hosking and Wallis (1997), if a region has N watersheds, for basin i, D_i as the measure of discordancy is computable as the following equation [16]:

$$D_i = \frac{1}{3}\, N(u_i - \bar{u})^T\, A^{-1}\, (u_i - \bar{u}) \tag{6}$$

Where u_i is a vector containing the L-moment ratios for basin i, $\bar{u}$ is the unweighted regional average for u_i and A is the matrix of sums of squares and cross products. For each data set is determined critical value. If the D_i value regarding any station placed in the cluster is less than the critical value, the cluster is homogeneous, otherwise, the heterogeneous stations should be removed from the cluster.

2.7. Best-fit distribution

Statistical tests are required to confirm the appropriateness of the chosen distribution which provides a certain degree of confidence. A test based on the Monte Carlo simulation described by Hosking and Wallis (1993, 1997) is used for this aim [16, 50]. Five statistical distributions are used that include Gen. logistic (GLO), Gen. Extreme value (GEV), Gen. Normal (GNO), Pearson type III (PE3), and Gen. Pareto (GPA). The goodness-of-fit measure for each distribution is given by:

$$Z^{DIST} = (\,\tau_4^{DIST} - \tau_4^R + B_4)/\sigma_4 \tag{7}$$

Where τ_4^{DIST} is the theoretical L–Kurtosis coefficient of the candidate probability distribution. τ_4^R is the L-moment ratio and B_4 is the bias of τ_4^R *and* σ_4 is the standard deviation of τ_4^R. A distribution could be declared as fitting satisfactorily if $|Z^{DIST}| \leq 1.64$ [50].

2.8. Multiple regression method

The most commonly used relation between the flow statistics (represented here by the flood-quantile Q_T of return period T years) and the watershed characteristics (A, B, . . . M) is the power-form function [48]. The multiple regression model can be expressed in the following form:

$$Q_T = \alpha\, A^a\, B^b\, C^c \ldots M^m \tag{8}$$

Where α is the regression constant and a, b, c, . . ., m are regression coefficients described by regression analysis. linear regression resulting from the logarithms of the variables provides multiple regression. Final multiple regression was applied to estimate flood discharge for given frequencies and watershed characteristics in relation to the homogeneous regions. The multiple regression technique facilitates flood peak magnitude determination in ungauged locations by transferring flood characteristics from similar sites where measured data are available. The relation is presented by flood-frequency equations [52]. The regression equations try to connect the most impressive watershed characteristics (independent variables) to flood characteristics (dependent variables; Q2, Q5, ..., Q200).

3. Results and discussion

3.1. Determination of effective independent parameters

Due to the fact that the first stage of regional frequency analysis is a close inspection of the data, statistical analysis seems essential. Gross errors and inconsistencies should be eliminated. Toward this aim, a check was accomplished, and based on the results the data were homogeneous (stationary) over time. All 17 sub-basins of the Gorganrood and Ghare-sou watersheds have been tested with different indices. Among them, one stations' data showed autocorrelation detected by the Autocorrelation Function Test (ACF). Additionally, three series failed Kendal's test. 13 sub-watersheds have remained for classification. 30 independent parameters were calculated and screened with PCA. All parameters are placed in 6 components that have similar effects on stream discharge. In each group, the most important one was selected as a component representative according to its variance (Red circles in Table 3). The results are arranged in Table 3. The six components allocated 91.87% of the total variance to themselves so all of them were used to determine effective variables.

3.2. Identification of homogeneous regions

Cluster analysis is a standard method of statistical multivariate analysis for dividing a data set into groups and has been frequently used to unify regions for regional analysis context. Regionalization methods such as cluster analysis need to select effective variables through similarity definition (or dissimilarity) for the watersheds [15]. Hosking and Wallis (1997) recommended methods that use watershed characteristics only for homogeneous region identification. Consequently, the use of watershed characteristics to conduct independent tests of the proposed regions is asked [16]. They believe Ward's method as a kind of hierarchical clustering method due to minimizing the Euclidean distance of characteristics space in each cluster is a powerful approach. In this study, Ward's clustering method was chosen for homogeneous regions determination (Fig. 2). At the next step, the L-moment test was accomplished to test the heterogeneity of the clusters.

The total area of 13 selected sites is 4784.663 km^2. The identified homogeneous regions (1) and (2) include 2673.81 and 302.4 km^2, respectively. All of the sub-basins except Sarmou located in region (1), and region (2) included just the Sarmou sub-basin. Among the 13 studied sub-watersheds, Sarmou sub-basin has the highest elevation values (minimum elevation of 500 meters and maximum elevation of 3600 meters). On the other hand, the mentioned area has the highest slope of the main stream, i.e. 10.92%. Due to its high elevation and slope magnitude, it has the lowest concentration time of 2.05 hours and has the highest flow speed of 3.85 m/s. The obvious difference between this sub-watershed and the other studied sub-watersheds in terms of topography and geology (possess the lowest percentage of permeable formations that is 0.25%) is the main reason for being heterogenous and forming a separate cluster

Table 3. The result of the Principal Component Analysis

Parameters	1st Component	2nd Component	3rd Component	4th Component	5th Component	6th Component
Min-e	-0.1	-0.24	0.13	-0.2	0.17	-0.08
Max-e	0.41	0.07	0.65	-0.41	-0.12	0.09
L	0.45	0.12	0.23	0.18	-0.33	-0.17
P	0.46	0.08	0.04	-0.10	-0.20	0.07
A	-0.47	-0.14	0.10	0.3	-0.03	-0.03
G-c	-0.44	0.36	-0.06	-0.55	0.54	0.35
S-L	-0.4	-0.28	0.18	0.02	0.09	-0.20
S-S	0.57	0.21	0.46	0.01	-0.08	0.15
Tc	-0.37	0.39	-0.05	-0.35	-0.17	-0.03
V	0.45	0.03	0.36	-0.42	-0.13	0.09
B-S	0.05	-0.41	0.36	-0.1	0.37	-0.30
P-A	0.52	-0.03	-0.34	0.55	-0.32	-0.31
N-A	0.4	0.09	-0. 27	-0.2	0.16	0.22
S-A	0.28	0.35	0.49	0.12	-0.08	-0.27
E-A	0.34	-0.13	-0.23	-0.01	-0.07	0.09
W-A	0.12	-0.45	0.78	-0.04	0.15	0.19
D-d1	-0.24	0.18	0.08	-0.10	0.56	0.24
D-d2	-0.36	0.23	0.12	-0.02	0.53	0.26
D-d3	-0.08	-0.18	0.08	-0.14	0. 51	-0.29
D-d4	-0.32	-0.2	0.52	-0.16	0.56	-0.07
TD-d	0.28	0.07	-0.38	0.15	-0.30	-0.64
P-F	0.19	0.19	0.33	-0.56	-0.06	-0.32
C-F	-0.43	-0.43	-0.42	0.31	0.21	0.02
F	-0.41	-0.41	0.19	0.15	-0.62	-0.19
Ra	0.12	0.12	-0.01	0.12	0.05	-0.41
Ag	-0.10	-0.10	-0.21	-0.34	-0.43	0.23
OT	0.4	0.4	0.24	0.49	0.1	-0.18
A-T	-0.17	-0.17	-0.93	0.21	0.11	-0.08
D-C	-0.13	-0.13	-0.65	-0.18	0.05	0.45
A-P	-0.02	-0.02	0.50	0.39	-0.31	0.27
Eigenvalue	3.79	2.37	1.91	1.55	0.85	0.56
% of variance	31.56	19.73	15.88	12.96	7.05	4.68
Cum. % of variance	31.56	51.29	67.19	80.13	87.19	91.87

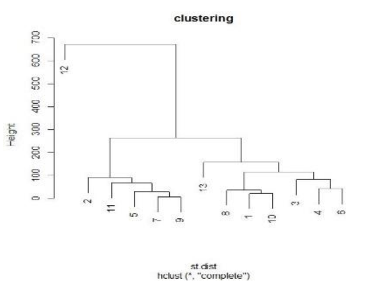
Min-e: minimum elevation, Max-e: maximum elevation, L: length, P: perimeter, A: area, G-co: Gravelius coefficient, S-L: stream's length, S-S: stream slope, Tc: concentration time, V: velocity of flow, B-S: basin's slope, P-A: area of the pediment, N-A: north land's area, S-A: south land's area, E-A: east land's area, W-A: west land's area, D-d1,2,3,4:1,2,3 and 4 ranked stream's density. TD-d: total density drainage, F: forest's area, Ra: ranches' area, Ag: agriculture field's area, OT: other land uses area, P-F: Percentage of permeable formations, C-F: carbonic bed rocks (namely percentage of impermeable formations), D-C: Demarten coefficient, A-T: annual average temperature, A-D: average annual discharge A-P: average annual precipitation.

3.3. Homogeneity of the clusters

As mentioned in the methodology section, L-moment is used to determine the homogeneity of groups. In the current study, discordancy values were measured and H was used to determine the heterogeneity. H-statistic is a statistical test based on L-moment ratios. H < 1 indicates that the region is acceptably homogeneous; 1 < H < 2 means possibly heterogeneous, and H >2 emphasizes definitely heterogeneous situations. The results show that the absolute values of the H-statistic for both clusters are less than 1, and these regions are acceptably homogeneous. The discordancy values regarding study clusters are presented in Table 4.

3.4. Identification of the best-fit distribution

This section includes the results of parameter estimation and the best-fit distribution test. In the L-moments approach, the location, scale, and shape parameters of five commonly used distributions in regional flood frequency analysis are calculated by the averages of the L-moments and L-moment ratios in homogeneous regions. The Z-statistic as a goodness-of-fit index was used for the identification of regional distribution applied in each region. As it is obvious in Table 5 the GPA presents the best goodness-of-fit at regions (1) and (2).



1. Taghi-Abad
2. Galikesh
3. Lazoureh
4. Pole-Ordougah
5. Gonbad
6. Tangrah
7. Bagh Salian
8. Naharkhoran
9. Agh Qala
10. Siah-Ab
11. Tamar
12. Sarmou
13. Araz kouse

Figure 2. Dendrogram of clustered watersheds using Ward's method

Table 4. The discordancy values regarding study clusters

Cluster	Sub-watersheds	Discordancy Coefficient	Threshold limit of heterogeneity coefficient	H-Statictic
1	Galikesh	1.01	3.00	0.98
	Tamar	1.30		
	Gonbad	1.10		
	Bagh Salian	0.80		
	Agh Qala	1.20		
	Araz kouse	0.89		
	Naharkhoran	1.09		
	Taghi-Abad	0.94		
	Siah-Ab	1.78		
	Lazoureh	0.85		
	Pole-Ordougah	1.00		
	Tangrah	1.00		
2	Sarmou		1.33	0.73

The Z value of this distribution is the lowest value of the Z-statistic (less than 1.64), as a result, the GPA could be introduced as the best-fit distribution for the flood analysis at the region (1) and (2). The values of the Z-statistic indicate that the GPA for the regions (1) and (2) gives close fits to the L-moments' regional average. The results are shown in Table 5.

Table 5. Goodness-of-fit analysis (ZDIST) for five different frequency distributions in the homogeneous regions.

Distribution	Homogenous region	
	Cluster 1	Cluster 2
GLO	0.25	1.51
GEV	-1.55	0.68
GPA	-5.39	-1.1
GNO	-1.63	0.64
PE3	-2.04	0.45

3.5. Multiple-regression results

The aim is to achieve a relationship between the dependent variable (QT) and independent variables (watershed characteristics) in each homogeneous region. The parameters of GPA distribution i.e. location (n), scale (a), and shape (k) were calculated for each homogeneous region by direct use of the data. Mentioned parameters are essential for QT estimation. The predicted flood values of the 2, 5, 10, 25, 50, 100, and 200 years were calculated for each site. The results are presented in Tables 6 and 7.

Table 6. The predicted flood magnitudes (m^3/s) by GPA in the cluster (1)

Sub-watershed	Return period (T), year						
	2	5	10	25	50	100	200
Sarmou	1.134	1.483	1.660	1.818	1.899	1.956	1.996

Table 7. The predicted flood magnitudes (m^3/s) by GPA in the cluster (2)

Sub- watersheds	Return Period (T), year						
	2	5	10	25	50	100	200
Tangrah	1.79	2.69	3.15	3.55	3.76	3.91	4.01
Tamar	1.54	2.31	2.70	3.05	3.23	3.36	3.45
Galikesh	2.29	3.45	4.03	4.55	4.82	5.00	5.14
Gonbad	6.81	10.23	11.96	13.50	14.29	14.84	15.24
Lazoureh	1.93	2.90	3.40	3.84	4.06	4.22	4.33
Araz kouse	5.05	7.59	8.87	10.02	10.60	11.02	11.31
Bagh Salian	2.78	4.17	4.88	5.51	5.83	6.06	6.22
Taghi-Abad	0.39	0.59	0.69	0.78	0.83	0.86	0.89
Agh Qala	10.87	16.31	19.08	21.54	22.79	23.68	24.31
Naharkhoran	0.30	0.45	0.52	0.59	0.62	0.653	0.67
Siah-Ab	1.88	2.82	3.30	3.73	3.94	4.10	4.21
Pole-Ordougah	0.45	0.68	0.79	0.89	0.95	0.98	1.01

The relation among discharges of selected recurrence intervals of the basin and independent parameters is determined by multiple-regression method. Output variables from PCA are entered into the multivariate regression model, but only several variables will be entered in the text of the final model. The coefficient related to each variable expresses the intensity of its effect on the discharge magnitude of the corresponding cluster. The most common method of estimating the regression model is the stepwise method (53). In this method, first, the variable that has the highest normal correlation coefficient with the dependent variable, if its F statistic is greater than the predetermined F (F entry), is entered into the equation. The next variables are also entered in the model in the same way, but simultaneously with the entry of each new variable, all the variables in the equation are examined (their partial F is calculated) and if each of them has lost its significance level, they will be removed from the process. This operation is repeated and at the end of it, no variable with a significant level lower than the determined level will be present in the equation. Mentioned process started with A-T (Annual Temperature) with 0.93 of negative correlation. It was followed by TD_d (Total drainage density), F (Forest's area), S-S (Stream Slope), P-F (Permeable Formation percentage), and C-F (Carbonic Formation percentage). In the following, to check the final formula of regional discharge estimation, for validation, regression was first performed in 10 stations. According to the results, the value of significance F for the measured flow rate and six independent variables is 0.03. Considering the confidence factor of 95% and the low value of Significance F (less than 0.05), the regression is acceptable. According to the Table 8, the P values for three independent variables were lower or very close to 0.05, so these three independent variables determine the major part of the flow rate changes.

Table 8. The results of regression among six independent variables and discharge values in 10 sub-watersheds.

	R^2	R^{2adj}		F Significance
Regression	0.96	0.88		0.032
Variables	**S-S**	**P-F**	**A-T**	**Intersection**
P-Value	0.049	0.007	0.064	0.050
Coefficient	-0.859	0.16	-1.08	20.98

The resulting relation will be as follows:

$$Q_T = \alpha(S - S) + \beta(P - F) + \gamma(\mathrm{A} - \mathrm{T}) + \mathrm{z} \qquad (9)$$

Where Q_T is the discharge magnitude for T-year return period in m^3/s. S-S, P-F, and A-T are stream slope, permeable formation percentage, and average annual temperature, respectively. The coefficients of the regression equation are also given in Table 8. The slope of the main stream has a negative effect on the magnitude of discharge. For a certain amount of precipitation, the slope of the stream will cause the quick exit of the discharge. Therefore, the retention of rain in the sub-watershed is short and the discharge will not affect the following days much. Therefore, the steep slope of the stream will decrease the average annual discharge. On the other hand, the percentage of permeable formations has been introduced as the second influencing variable with a positive effect on the flow rate. The high infiltration of rainfall and runoff in the watershed slows down and reduces the amount of water reaching the river and finally to the outlet of the watershed. But in the watersheds near the sea, due to the saturation of the aquifers, sometimes it is not possible for water to penetrate. Therefore, the precipitation flows quickly and turns into discharge. In the studied watersheds, the rate of water infiltration in impermeable carbonate formations is much higher than in other formations due to deep and large fractures. The third variable affecting the discharge rate is the average annual temperature with a negative impress.

In the next step the final relation was validated in three sub-watersheds that were not included in the regression. The results are summarized in Table 9.

Table 9. The results of equation validation in three sub-watersheds (Gonbad, Naharkhoran and Lazoureh)

Sub-watershed	Observed (m^3/s)	Estimated (m^3/s)
Gonbad	2.09	3.05
Naharkhoran	0.43	0.44
Lazoureh	2.48	2.21

Based on the validation estimations, the regional model could be acceptable.

4. Conclusion

The present study reports a regional analysis carried out in the north of Iran, with the aim of L-moments approach evaluation for predicting flood discharge at different return periods. An accurate data screen of about 13 sub-watershed gauging stations was accomplished with the help of existing flood data for regional flood frequency analysis. Firstly, the whole study area was analyzed and then was categorized as two sub-regions defined by cluster analysis technique. In conditions with short data records, the application of probability distributions for predicting the flood magnitude in different return periods does not seem rational. However, the application of the L-moments technique facilitates increasing the data length at regional flood frequency analysis. The L-moments approach simultaneously uses several homogeneous watershed data during a hydrologic analysis.

The results of the PCA technique for the determination of the main variables showed that the 30 independent variables could be summarized into six components. The mean annual temperature, total drainage density, forest's area, stream slope, permeable formation percentage, and carbonic formation percentage of sub-watersheds were identified as the most important variables of the six components. Mentioned variables as representatives of six components carry the variation of discharge magnitude (91.87% of total variance).

The Ward's method as a hierarchical clustering method is a proper approach for regionalization in hydrology studies. The results of homogeneity test based on L-moments approach at the whole study area indicated the study area should be divided into two homogeneous sub-regions. The least number of series to calculate regional equations is 5 and because of inadequate data series having an equation for the region (2) with one sub-basin was not possible. So the whole case study has been considered as one homogeneous region. The results of the L-moment ratios and the Z-statistic criteria identified GPA distribution as the fittest distribution among five candidates for all the proposed clusters of the study area.

Finally, based on the results of clustering and stepwise regression, only three variables defining flood discharge magnitude were considered. In general, it can be said that due to the fact that the synoptic stations with more than 30 years of data series as well as reliable statistical tests have been used for data analysis, the regional estimation of discharge with the used methods can be accepted with great confidence. This regional model can be used to estimate discharge in areas with similar climatic and hydrological conditions where measured data are not available.

References

[1] Hamed, K. and Rao, A.R. eds. 2019. Flood frequency analysis. CRC press.

[2] Lawrence, D. 2020. The uncertainty introduced by flood frequency analysis in projections for changes in flood magnitudes under a future climate in Norway, Journal of Hydrology: Regional Studies, 28, p.100675.

[3] Hasan, I.F. 2020. Flood Frequency Analysis of Annual Maximum Streamflows at Selected Rivers in Iraq, Jordan Journal of Civil Engineering, 14.4

[4] Malekinezhad, H., Nachtnebel, H.P. and Klik, A. 2011. Comparing the index-flood and multiple-regression methods using L-moments, Physics and Chemistry of the Earth, Parts A/B/C, 36.1-4, p.54-60.

[5] Maghsood, F.F., Moradi, H., Massah Bavani, A.R., Panahi, M., Berndtsson, R. and Hashemi, H. 2019. Climate change impact on flood frequency and source area in northern Iran under CMIP5 scenarios, Water, 11.2, p. 273.

[6] State Water Policy, Government of Rajasthan, 1999. Department of Irrigation. http://waterresources.rajasthan.gov.in/.

[7] Stedinger, J.R. and Tasker, G.D. 1985. Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared, Water Resources Research, 21.9, p.1421-1432.

[8] GREHY, G.D.R.E.S. 1996. Presentation and review of some methods for regional flood frequency analysis, Journal of hydrology (Amsterdam), 186.1-4, p.63-84.

[9] Jingyi, Z. and Hall, M.J. 2004. Regional flood frequency analysis for the Gan-Ming River basin in China, Journal of Hydrology, 296.1-4, p.98-117.

[10] Bhat, M.S., Alam, A., Ahmad, B., Kotlia, B.S., Farooq, H., Taloor, A.K. and Ahmad, S. 2019. Flood frequency analysis of river Jhelum in Kashmir basin, Quaternary International, 507, p.288-294.

[11] Cafiero, L., Monforte, I., Mazzoglio, P., Ganora, D., Laio, F., Claps, P. and Viglione, A. 2023. Bayesian Spatially Smooth Regional Estimation of flood quantiles: Case study in Northern Italy, In Titolo volume non avvalorato, IUGG.

[12] Kim, N.W., Lee, J.Y., Park, D.H. and Kim, T.W. 2019. Evaluation of future flood risk according to RCP scenarios using a regional flood frequency analysis for ungauged watersheds, Water, 11.5, p.992.

[13] Acreman, M. and Wiltshire, S. 1989. The regions are dead. Long live the regions. Methods of identifying and dispensing with regions for flood frequency analysis, IAHS-AISH publication, 187, p.175-188.

[14] Burn, D. H. 1990. Evaluation of regional flood frequency analysis with a region of influence approach, Water Resources Research, 26.10, p.2257-2265.

[15] Burn, D.H. 1997. Catchment similarity for regional flood frequency analysis using seasonality measures. Journal of hydrology, 202.1-4, p.212-230.

[16] Hosking, J.R.M. and Wallis, J.R. 1997. Regional frequency analysis, p. 240.

[17] Shu, C. and Burn, D.H. 2004. Homogeneous pooling group delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement. Journal of Hydrology, 291.1-2, p.132-149.

[18] Ouarda, T.B.M.J., Ba, K., Diaz-Delgado, C., Carsteanu, A., Gingras, H., Quentin, E., Trujillo, E. and Bobee, B. 2007, May. Regional flood frequency estimation at ungauged sites in the Balsas River Basin, Mexico, In AGU Spring Meeting Abstracts (Vol. 2007, p. H51B-02).

[19] Shu, C. and Ouarda, T.B. 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system, Journal of Hydrology, 349.1-2, p.31-43.

[20] Pagliero, L., Bouraoui, F., Diels, J., Willems, P. and McIntyre, N. 2019. Investigating regionalization techniques for large-scale hydrological modelling, Journal of hydrology, 570, p.220-235.

[21] Song, Z., Xia, J., Wang, G., She, D., Hu, C. and Hong, S. 2022. Regionalization of hydrological model parameters using gradient boosting machine, Hydrology and Earth System Sciences, 26.2, p.505-524.

[22] Tasker, G.D. and Thomas Jr, W.O. 1978. Flood-frequency analyses with prerecord information, Journal of the Hydraulics Division, 104.2, p.249-259.

[23] Acreman, M.C. and Sinclair, C.D. 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland, Journal of Hydrology, 84.3-4, p.365-380.

[24] Stamey, T.C. and Hess, G.W. 1993. Techniques for estimating magnitude and frequency of floods in rural basins of Georgia, Water-Resources Investigations Report, 93, p.4016.

[25] Kjeldsen, T.R., Smithers, J.C. and Schulze, R.E. 2002. Regional flood frequency analysis in the KwaZulu-Natal province, South Africa, using the index-flood method, Journal of hydrology, 255.1-4, p.194-211.

[26] Farsadnia, F., Kamrood, M.R., Nia, A.M., Modarres, R., Bray, M.T., Han, D. and Sadatinejad, J. 2014. Identification of homogeneous regions for regionalization of watersheds by two-level self-organizing feature maps, Journal of Hydrology, 509, p.387-397.

[27] Calegario, A.T., Pruski, F.F., Ribeiro, R.B., Ramos, M.C. and Rego, F.S. 2020. Physical analysis of regionalized flow as an aid in the identification of hydrologically homogeneous regions, Engenharia Agrícola, 40, p.334-343.

[28] Walling, D.E. and Fang, D. 2003. Recent trends in the suspended sediment loads of the world's rivers, Global and planetary change, 39.1-2, p.111-126.

[29] Walling, D.E. 2006. Human impact on land–ocean sediment transfer by the world's rivers, Geomorphology, 79.3-4, p.192-216.

[30] Lamontagne, J.R., Stedinger, J.R., Yu, X., Whealton, C.A. and Xu, Z. 2016. Robust flood frequency analysis: Performance of EMA with multiple Grubbs-Beck outlier tests, Water Resources Research, 52.4, p.3068-3084.

[31] Khazar Water Consulting Engineers Company, 1377. Report on the integration of studies on water resources of the Gorganrood and Qarasu basins, vol. 3.

[32] Jamab Consulting Engineers, 1378, Report on Country Water Master Plan, Total and Coastal Watershed, Ministry of Energy, Tehran.

[33] Nautiyal, M.D. 1994. Morphometric analysis of a drainage basin using aerial photographs: a case study of Khairkuli Basin, District Dehradun, UP, Journal of the Indian Society of Remote Sensing, 22, p.251-261.

[34] Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components, Journal of educational psychology, 24.6, p.417.

[35] Zavareh, M., Maggioni, V., and Sokolov, V. 2021. Investigating water quality data using principal component analysis and granger causality, Water, 13.3, p.343.

[36] Erunova, M.G., and Yakubailik, O.E. 2023. Methods and technologies for spatial analysis of regional ecosystems based on the watershed approach, Integrated Environmental Assessment and Management, 19.4, p.972-979.

[37] McKenna Jr, J.E. 2003. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis, Environmental Modelling & Software, 18.3, p.205-220.

[38] Mohammadi, S.A. and Prasanna, B.M. 2003. Analysis of genetic diversity in crop plants—salient statistical tools and considerations, Crop science, 43.4, p.1235-1248.

[39] Vega, M., Pardo, R., Barrado, E. and Debán, L. 1998. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis, Water research, 32.12., p.3581-3592.

[40] Otto, M. 1998. Multivariate methods. Analytical chemistry, 916.

[41] Cupak, A. and Michalec, B. 2022. Regionalisation of watersheds with respect to low flow, Journal of Water and Land Development, 55.

[42] Ward Jr, J.H. 1963. Hierarchical grouping to optimize an objective function, Journal of the American statistical association, 58.301, p.236-244.

[43] Willmott, C.J. and Vernon, M.T. 1980. Solar climates of the conterminous United States: A preliminary investigation, Solar Energy, 24.3, p.295-303.

[44] Winkler, J.A. 1985. Regionalization of the diurnal distribution of summertime heavy precipitation. In Preprints, Sixth Conf. on Hydrometeorology, Indianapolis, IN, Amer. Meteor. p. 9-16.

[45] Kalkstein, L.S. and Corrigan, P. 1986. A synoptic climatological approach for geographical analysis: assessment of sulfur dioxide concentrations, Annals of the Association of American Geographers, 76.3, p.381-395.

[46] Nathan, R.J. and McMahon, T.A. 1990. Identification of homogeneous regions for the purposes of regionalization, Journal of Hydrology, 121.1-4, p.217-238.

[47] Adhami, M. 2012. Estimation of suspended sediment load using physical characteristics in Gorganroud and Ghareh-Sou Watersheds. Gorgan University of Agricultural Sciences and Natural Resources. MS Thesis, p125, Gorgan, Iran.

[48] Hosking, J.R.,1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics, Journal of the Royal Statistical Society Series B: Statistical Methodology, 52.1, p.105-124.

[49] Vogel, R.M. and Fennessey, N.M. 1993. L moment diagrams should replace product moment diagrams, Water resources research, 29.6., p.1745-1752.

[50] Hosking, J.R.M. and Wallis, J.R. 1993. Some statistics useful in regional frequency analysis, Water resources research, 29.2., p.271-281.

[51] Pandey, G.R. and Nguyen, V.T.V. 1999. A comparative study of regression based methods in regional flood frequency analysis, Journal of Hydrology, 225.1-2, p.92-101.

[52] Yu, G., Wright, D.B., Zhu, Z., Smith, C. and Holman, K.D. 2019. Process-based flood frequency analysis in an agricultural watershed exhibiting nonstationary flood seasonality, Hydrology and Earth System Sciences, 23.5, p.2225-2243.

[53] Moazzami, M., Feyznia, S. 2016. Regional Analysis of Suspended Sediment (Case Study of Jarhiri River), 4th National Conference on Watershed Science and Engineering of Iran, Watershed Management, Karaj.