Partitioning around medoids approach application for computation of regional flood and landslide quantiles

Gnanaprakkasam Shiyamalagowri[1,*], *Pattukandan* Ganapathy[1], *Vladislav* Zaalishvili[2,3], *Dmitry* Melkov[2]

[1]Centre for Disaster Mitigation and Management, Vellore Institute of Technology, 632014 Vellore, India
[2]Geophysical Institute, Vladikavkaz Scientific Centre, Russian Academy of Sciences, 362002 Vladikavkaz, Russia
[3] North-Ossetian State University named after K.L. Khetagurov, 362025 Vladikavkaz, Russia

Abstract. Flood and landslides causes serious damage to the functioning of the society which results in a huge loss of human life, material and other environmental impacts. In this paper, partitioning around medoids approach is executed for the assessment of flood quantiles over 145 sites using 11 basin characteristics. The study region is classified into 6 clusters as a result of the partitioning algorithm which are further proved to be homogeneous by applying the heterogeneity measure test. Results from the study provided the regional flood quantile measurements for the ungauged sites derived from L moments with good accuracy limits for the recurrence intervals 50,100,200 and 500 years. As a floods landslides may caused by rainfalls, especially over long time periods, which both increase the weight of slopes and can lubricate planes of weakness within rock or sediment. It is shown that landslides are also allocated in some of the clustered zones, depending of geological conditions of the clusters. Thus regional flood quintiles in conjunction with geology and topography forms landslide activity quantiles.

1 Introduction

The problem of preventing and mitigating the severity of hazardous natural and anthropogenic processes effect is relevant for most of the world's territories, including Russia, India and . Expected economic and social risks are particularly high in the North Caucasus due to high population density and high level of volcanic, seismic, landslide, glacial, mudflow, flood and other hazards. At the same time, specific factors of the evolution of the natural environment can provoke catastrophic natural and natural-anthropogenic processes that were not previously studied at a sufficient level for a reliable prognosis and, therefore, often unexpected and destructive: a powerful landslide near the village of Mizur in 2002 and surface failures in the area of village Sadon, as a response to the intensive and irrational mining activity in many respects in the past. In 2002, Kolka glacier unexpectedly

* Corresponding author: seismogans@yahoo.com

collapsed along the northern slope of the Kazbek volcano in the Karmadon Gorge, which caused the death of more than 100 people and caused an ecological catastrophe [1]. Later, on May 17, 2014, from the opposite slope of Kazbek, a large amount of rocks and ice collapsed in the area of the Devdorak glacier in the territory of Georgia. One of the main factors assumed is high rainfall preceding the mentioned catastrophic events. Water saturation is critical in such processes as liquefaction phenomenon [2, 3]. In August 2018 a large number of people died (more than 400 people) as a result of landslides and downpours in the south and west of India. Flooding and landslides have led to significant violations of rail and road traffic in the south. In India, the current monsoon rainy season is one of the most active in the last 100 years.

The most dangerous hydrological natural hazard is a riverine flood which is caused by the overflow of water from the river channel into the dry surface. The influences of flood are very high as it endangers the life of human being and especially the engineering constructions near the rivers. Flood frequency analysis determines the relationship between the magnitude and frequency of floods which are essential in constructing the dams, culverts and bridges [4]. The at-site estimates of flood can be determined easily, since the data are readily available at the gauge [5]. But the prediction of flood quantile magnitude for ungauged sites is one of the biggest challenges faced by water resource engineers, where the amount of information available is very less [6-9]. Regional flood frequency analysis (RFFA) evaluates the information about the flood at the ungauged sites which is useful in flood plain management. The important steps in RFFA are to identify the homogeneous regions, choosing the appropriate frequency distribution with respect to regionalization and the final step is to determine the flood quantiles at the target site [10-12]. Annual maximum flood (AMF) and peaks over threshold (POT) are widely used in the prediction of extreme events. So many studies have been carried out to develop the regionalization approaches which were designed for the assessment of extreme events in the ungauged sites by extracting information from the similar group of gauged watersheds [13-14]. Region of influence (ROI) approach applies Euclidean distance to group the sites based on the information extracted from the gauged sites with respective to the target site considered [15].

Clustering techniques are widely used in regionalization where the similar sites are placed under one cluster and dissimilar sites occupy the other group using partitioning and hierarchical clustering algorithms [16]. Fuzzy based clustering is combined with self-organizing feature map to delineate the watersheds for estimating the flood quantiles in the ungauged sites [17]. A new regional flood frequency analysis approach based on anarchy connected with each and every data set for detecting the outliers was developed in recent years [18]. Canonical correlations analysis mainly concentrates on the linear relationship between the variables when applied to the ungauged streams [19]. Partitioning Around Medoids (PAM) clustering algorithm deals with distance matrix that focus on medoids as the cluster centers which is very useful in the case when the data points cannot be assigned to any of the cluster [20].

Many studies have concentrated on delineating homogeneous regions and in estimation of regional quantiles using non stationarity principles [21–23]. This study aims at grouping the sites based on medoids rather than traditional k-means method that uses means as the cluster centroid which suffers from major drawbacks in dealing with noise and outliers. Also, the medoids based approach concentrates in reducing the sum of dissimilarities between the pairs of data points rather than considering the squared sum of Euclidean distances as in the case of k-means. The objectives of the study are 1) to classify the region into groups by applying the partitioning around medoids algorithm 2) to determine the homogeneity of the clusters formed 3) to trace out the best frequency distribution that fits the data 4) to estimate the flood quantiles for those homogeneous sites and also to derive the parameters for the

distributions at 0.90 level of significance 5) to evaluate the performance of the approach using the error statistics.

2 Study area

Partitioning around medoids approach for the assessment of flood quantiles is executed over 145 sites in Indiana watershed which is in the east-central United States (Figure 1). The annual maximum peak values for all the stream flows sites with more than 10 years of data were obtained from United States Geological Survey (USGS). The physiographic attributes are drainage area, main channel slope, main channel length, average basin elevation, storage (percentage of the drainage area contributed to the study which is covered by water bodies), forest (percentage of the drainage area enclosed by forest), soil runoff coefficient and location of sites. The meteorological attributes are mean annual precipitation and I24,2 (24-hour rainfall having a return period of 2 years). The information regarding the attributes were extracted from [24]. While, the cluster analysis is applied extensively to the attributes of 145 sites and totally the study region is divided into 6 clusters.

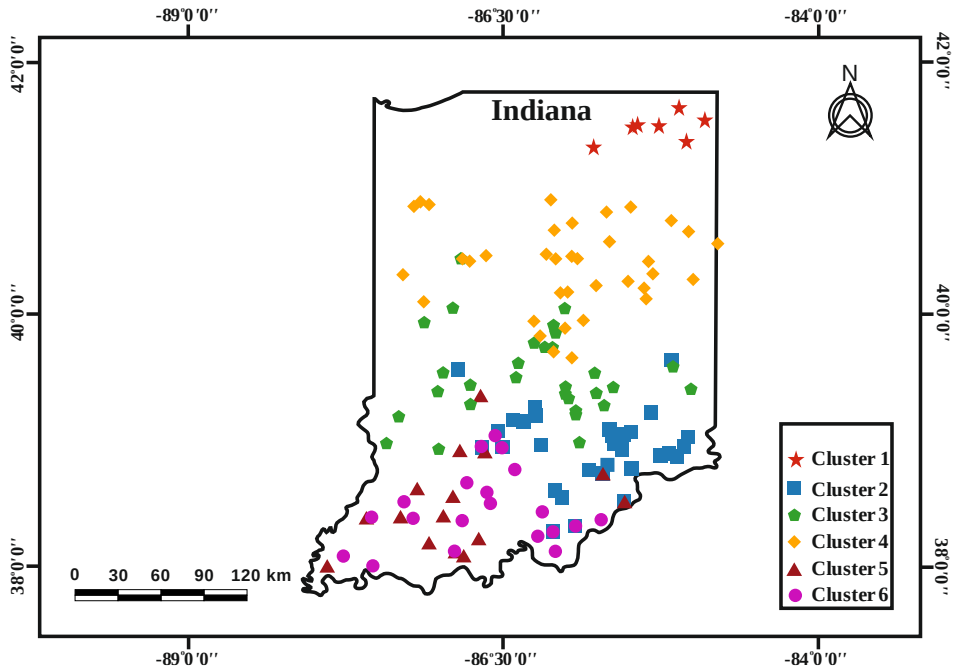

Fig. 1. Clusters formed based on partitioning around medoids clustering approach

3 Methodology

Discordancy measure is carried out to eliminate the conflicting sites from the group. If the discordancy value for any site exceeds the critical value 3, then those sites are removed from the group [25]. The shape of the frequency distribution is obtained using the L-moments. The location of the distribution, scale or dispersion, skewness and kurtosis of the distribution are also derived using the L-moments.

The cluster analysis aims at partitioning the data sets by following the principle, that similar data sets occupy one cluster and dissimilar data sets forms the other one. In this study, PAM (Partitioning around medoids) is used to cluster the Indiana water shed based on the

attributes as mentioned in previous section. Where, PAM's method is a partitioning based clustering that mainly concentrates on the medoids of the data points which are actually the centroids that are used to form the cluster [26-28]. Sequence of objects are first identified based on medoids and it forms a set T. Let Q be the set of objects and R is the set of unselected objects. Each and every data point is assigned two numbers with respect to the dissimilarity between the object to the nearest object in the total set T (Ki) and the dissimilarity between the object to the second nearest object in the total set T (Li). During the SWITCH process, these two numbers must be updated regularly.

Homogeneity measure is applied to the clusters formed in order to find out the homogeneous groups. The observed variations are compared to the homogeneous regions using the L moment ratios for all the sites.

Once the clusters are tested for homogeneity, the next stage is to select the suitable distribution for each and every cluster group that would produce precise flood quantiles estimation. This can be achieved by L-Moment ratio diagrams in which the L-Skewness and L-Kurtosis values are compared. The best distribution is selected, when most of the data points lie closer to that line, where three parameters of the 5 distributions namely generalized logistic distribution (GLO), generalized extreme value distribution (GEV), generalized pareto distribution (GPA), generalized normal distribution (GNO), and Pearson Type III distribution (PE3) are plotted in a graph. To precisely select the distribution better than the graphical representation, goodness of fit test is applied to the data points, which is given by Z_{dist} value in Equation (1).

$$Z_{dst} = \frac{(K_{dist} - K^R + B^R)}{\sigma_R} \tag{1}$$

where dist value denotes either of the 5 distributions, K_{dist} gives L-Kurtosis of the distribution, K^R represents regional average L-Kurtosis which weights uniformly to the record length of the sites, B^R is the bias value of the sample L-Kurtosis and σ_R is the standard deviation. If the Z_{dist} value is closer to 0, then it is considered as a satisfactory measure. If $Z_{dist} \leq 1.64$, the distribution is said to lie within the critical level. Development of the regional flood frequency relationships is essential in estimating the quantiles of flood corresponding to recurrence intervals 50,100,200 and 500 years. Each site quantiles are obtained by using the parameter M which represents the vector of probabilities. Where, M (0.98) denotes 50 years, M (0.99) denotes 100 years, M (0.995) denotes 200 years and M (0.998) denotes 500 years of return period. Regional estimates are derived by considering one site at a time as ungauged and treating all other sites in the group as gauged. Each site's quantile function can be obtained by multiplying the mean of each site by its regional growth curve. Growth curve qi for site i is given by Equation (2).

$$\mathrm{q(M)} = \frac{\mathrm{Qi(M)}}{\mu_i} \tag{2}$$

Where Qi is the site's quantile function and the mean for site i is given by μi which is also the measure of index flood. Thus, the estimation of floods for each and every site is derived by the product of normalized regional quantile and first order of L-moments, i.e. mean of each site. Identical results are assumed to be estimated for clusters with same frequency distribution. To estimate regional flood quantile $\widehat{Q}p^R$ for each site i, with respect to a T year return period, the Equation (3) is used.

$$\widehat{\mathrm{Q}}\mathrm{p}^{\mathrm{R}} = \frac{\sum_{i=1}^{n} l_i \widehat{\mathrm{Q}}\mathrm{p}^{(i)}}{\sum_{i=1}^{n} l_i} \tag{3}$$

4 Results and discussion

In this study, out of 145 sites, 3 sites were found to be greater than the critical value 3 according to the discordance measure as discussed in the previous section. Site number 3364570 varies distinctly from the other sites with the discordancy value 7.47 which is consider as very higher than the critical value 3. Whereas the discordancy values of the site numbers 3342180 and 3335700 are slightly greater than the critical value occupying the border level. Hence those 3 sites are removed from the group and remaining 142 sites are considered for the cluster analysis. Table 1 gives the information about the 3 discordant sites along with their L moment ratios which clearly illustrates the site number, size, mean of annual max flood, L-CV (co-efficient of variation), L-Skewness, L-Kurtosis and the discordancy measure of the discordant sites.

Table 1. Details of discordant sites with L-Moments

S.no	Site number	Size	Mean	Coefficient of Variation	L-Skewness	L-Kurtosis	Discordancy Measure
1	3364570	10	149.1	0.3721	0.0015	-0.1756	7.47 **
2	3342180	10	180	0.4119	0.3579	0.2459	3.80 *
3	3335700	32	5764.062	0.2451	0.0121	0.1179	3.24 *

4.1 Formation of Homogeneous Groups

PAM clustering approach is applied to the 142 sites of Indiana watershed in order to form a group of 6 clusters which are considered to be "Definitely Homogeneous". The heterogeneity measure test has been extensively worked out to derive the status of the cluster with respect to homogeneity. Table 2 illustrates the regional frequency distribution (RFD), status of homogeneity, H_m values, number of sites present in each cluster. All the clusters are computed to be homogeneous with the H_m values of -0.4072, 0.97910, 0.90992, 0.82837, 0.66277, 0.46641 respectively for each individual 6 clusters. Since all the H_m values lies within the threshold level 1, hence the entire region falls under the category of "Definitely Homogeneous". Fig. 1 visually shows the dispersion of all the sites according to their respective clusters, which were partitioned into and derived to be as homogeneous. This is the most difficult step in regional flood frequency analysis as it requires a more descriptive knowledge about all the site characteristics. The algorithm identifies the k representative objects which are considered to be the cluster centers with respect to their medoids. Silhouette width plays a major role in forming the regions. If the silhouette width is positive, then the sites exactly belongs to that cluster. If the silhouette width is negative, then the site belongs to the neighboring cluster. In the other case, if the silhouette width is 0, then the site is in the border of both the cluster. Based on the width, various adjustments are made to the sites so that they definitely form the homogeneous regions.

Table 2. Details of the clusters formed by partitioning around medoids approach

Clusters	Number of sites	H_m value	Homogeneity Status	RFD
1	7	-0.4072	Homogeneity	GEV
2	31	0.97910	Homogeneity	GEV
3	33	0.90992	Homogeneity	GEV
4	36	0.82837	Homogeneity	GLO
5	15	0.66277	Homogeneity	PE3
6	20	0.46641	Homogeneity	GEV

4.2 Flood frequency relationship

In Figure 2, most of the data points lie closer to the GEV distribution line for clusters 1,2,3 and 6, whereas sites are closer to GLO distribution line in the case of cluster 4 and finally PE3 proves to be the best distribution by attracting more sites towards the line in the case of cluster 3. Hence L-moment ratio diagrams pictorially represents the appropriate distribution by comparing the L-skewness and L-kurtosis values of the sites chosen in the cluster rather than the numerical values. Regional flood frequency relationship is developed to estimate the quantiles of flood with the vector of probabilities 0.98,0.99,0.995 and 0.998 for the return periods 50, 100, 200 and 500 years respectively.

The product of mean with the regional growth curve gives the quantile functions for every site whereas regional estimates are derived by considering one site at a time as ungauged and combining all the sites together and substituting those values in Equation (2), regional growth curve are framed. At site and regional quantiles are compared using GEV, GLO and PE3 distributions.

The relationship between the at-site and regional quantiles using the GEV, GLO and PE3 distributions for the return periods T= (50 years, 100 years, 200 years and 500 years) implies that the data points lie closer to the solid line in 1:1 ratio. Medoids based approach classifies the streams in Indiana into 6 clusters and L-moment approach is applied to the analysis for quantile derivation. According to the performance ratings of each metrics this study had provided the best results where, NSE, d, r and R^2 values are closer to 1 according to the criterion.

5 Results and Discussion

Regional flood frequency analysis using the PAM clustering approach was employed on 145 sites in the study area and finally classified into 6 homogeneous clusters, where all the H values were derived to lie below 0. Goodness of fit measure was employed on the data sets to find the best distribution that suits to estimate the flood quantiles where, Z statistics had identified GEV as the best suitable distribution for the clusters 1, 2, 3 and 6 with the values closer to zero. And for the clusters 4 and 5, the best distributions selected were GLO and PE3.

The performance of the approach is validated and hence the best results are provided and proved to be more accurate by the metrics NSE, index of agreement, Pearson correlation coefficient and coefficient of determination with the values closer to 1. Based on the results, medoids based clustering approach proved to be the robust method when compared with the traditional k-means based clustering which suffers from drawbacks in dealing with noise and outliers. The regional flood quantiles generated for the ungauged sites were derived from L moments with good precision bounds for the recurrence intervals 50,100,200 and 500 years. Thus, the computation of flood seems to be very important for the civil engineering structures constructed at or near the waterbodies whose rarity is hence determined by the various return periods, which is allied with flood control.

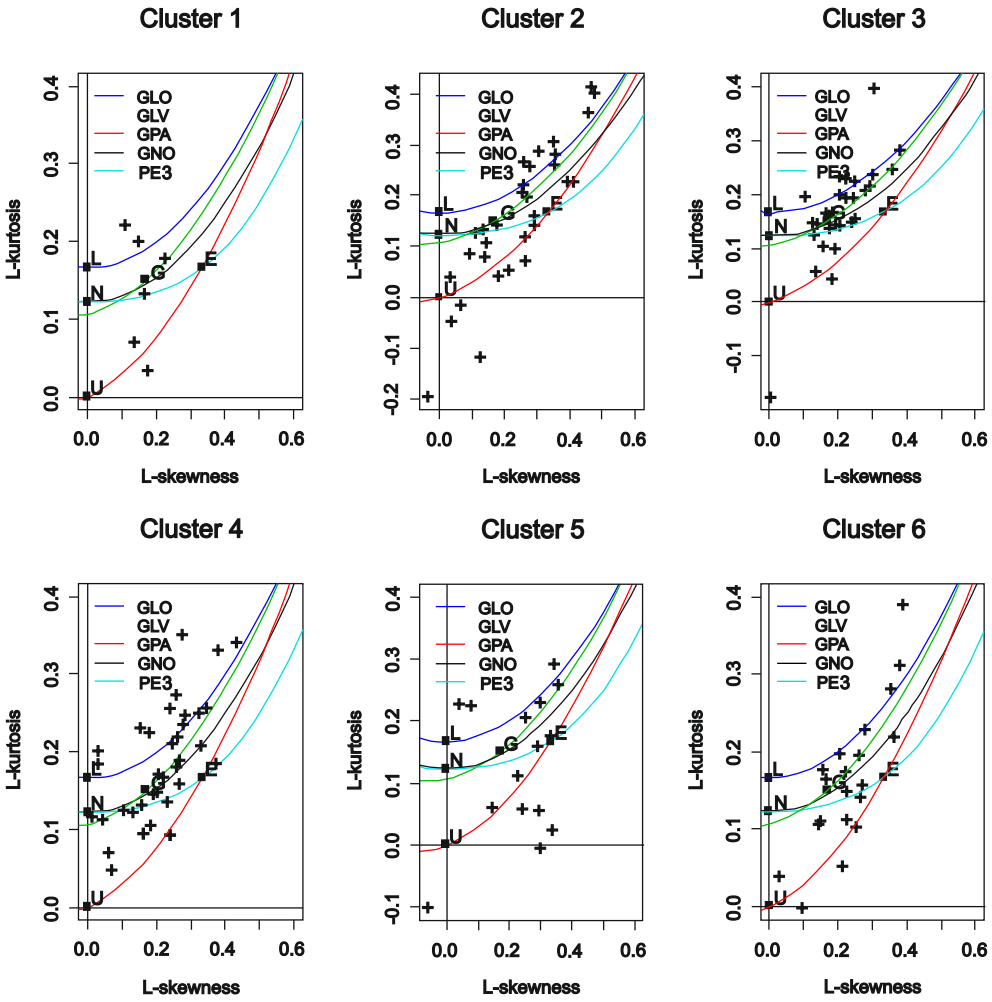

Fig. 2. L-moment ratio diagrams for the 6 clusters

Landslide processes strongly depends of hydrological factors. Data on historical landslides on the investigated territory can be found in [29], information on current landslides is presented in the Indiana Standard Hazard Mitigation Plan (SHMP, Indiana Department of Homeland Security, Polis Center of Indiana University, 2019). Compiled map in fig. 3 shows that landslides are also trends to have a cluster structure. Much of them are concentrated in cluster 5 and cluster 6 intersection the next huge amount of landslide processes are in the region of cluster 2 and 3, which are of Mississippian, Ordovician and Sillurian bedrock geology. Other events may be considered as background ones. Thus landslides are correlates with flood quantiles, depending on geology and topography. Application of modern complementary approaches, including considered one in integration with the methods of system analysis leads to multifactor methodology for exogenous geological events estimation and development of an effective system for mitigation.

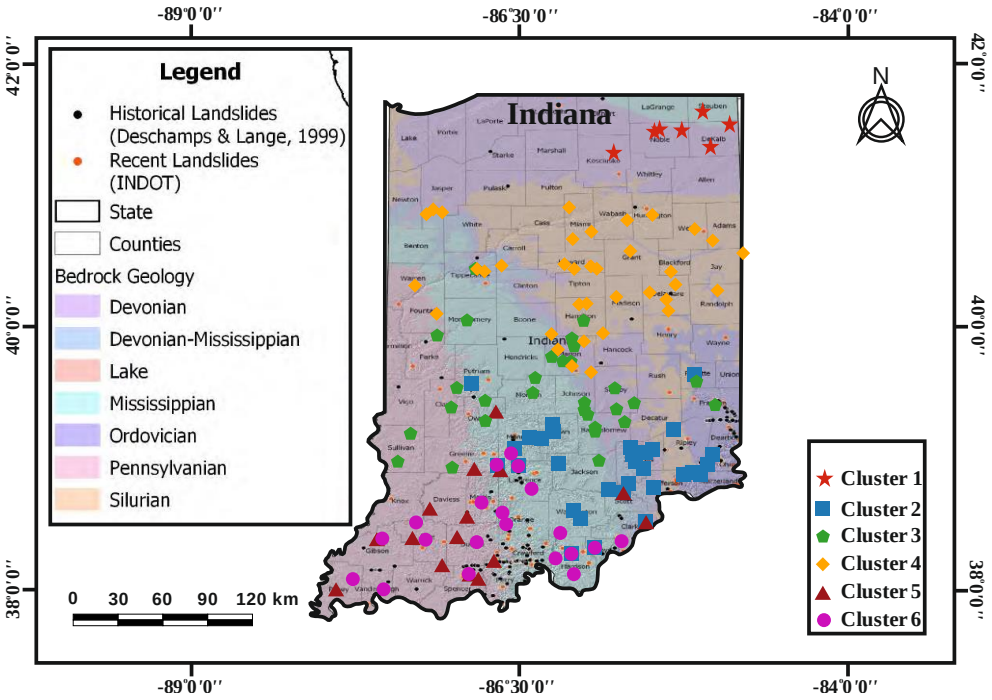

Fig. 3. Historical Lamdslides, bedrock geology compared with clusters

Acknowledgement:

The research was supported by Russian Science Foundation (Project No. 19-47-02010 RSF -DST (2018):" Natural hazards and monitoring for mountain territories in Russia and India". The authors are grateful to the Centre for Disaster Mitigation and Management (CDMM), Vellore Institute of Technology (VIT), Vellore for providing their constant support with continuous encouragement in conducting the study and also would like to express sincere thanks to Dr. Roshan Srivastav, former Associate Professor, Centre for Disaster Mitigation and Management, Vellore Institute of Technology, Vellore for providing his guidance.

References:

1. V. Zaalishvili, D. Melkov, Izvestiya, Physics of the Solid Earth, **50** (2014). doi: 10.1134/S1069351314050097
2. G. Ganapathy, V. Zaalishvili V.B., D. Melkov, B. Dzeranov, S. Chandrasekaran, Geology and Geophysics of Russian South, **8** (2018). doi: 10.23671/VNC.2018.3.16552
3. V. Svalova, V. Zaalishvili, G. Ganapathy, A. Nikolaev, D. Melkov, Geology and Geophysics of Russian South, **9** (2019). doi: 10.23671/VNC.2019.2.31981
4. V. Chow, D. Maidment, L. Mays *Applied Hydrology* (McGraw Hill, 1988). doi: 10.1080/02626666509493376
5. A. Rahman, A. Rahman, M. Zaman, K. Haddad, A. Ahsan, Nat Hazards, **69** (2013). doi: 10.1007/s11069-013-0775-y
6. J. Stedinger, Water Resour Res., **19** (1983). doi: 10.1029/WR019i002p00503

7. Z. Zrinji, D. Burn, J Hydrol., **153** (1994). doi: 10.1016/0022-1694(94)90184-8
8. T. Gado, V. Nguyen, J Hydrol Eng., **21** (2016). doi: 10.1061/(ASCE)HE.1943-5584.0001312
9. S. Grimaldi, A. Petroselli, F. Serinaldi, Hydrol Sci J., **57** (2012). doi: 10.1080/02626667.2012.702214
10. Grehys, J Hydrol., **186** (1996). doi: 10.1016/S0022-1694(96)03042-9
11. R. Kumar, C. Chatterjee, S. Kumar, A. Lohani, R. Singh, Water Resour Manag., **17** (2003). doi: 10.1023/a:1024770124523
12. K. Yarrakula, B. Samanta, D. Deb, Int J Hydrol Sci Technol., **5** (2015). doi: 10.1504/ijhst.2015.070097
13. V. Srinivas, ISH J Hydraul Eng., **15** (2009). doi: 10.1080/09715010.2009.10514974
14. M. Durocher, F. Chebana, T. Ouarda, Q. Trois-rivières, B. Forges, Hydrol Earth Syst Sci., **20** (2016). doi: 10.5194/hess-20-4717-2016
15. D. Burn, Hydrol Sci J., **35** (1990). doi: 10.1080/02626669009492415
16. A. Kar, N. Goel, A. Lohani, G. Roy, J Hydrol Eng., **17** (2012). doi: 10.1061/(asce)he.1943-5584.0000417
17. V. Srinivas, S. Tripathi, R. Govindaraju, A. Rao, J Hydrol., **348** (2008). doi: 10.1016/j.jhydrol.2007.09.046
18. B. Basu, V. Srinivas, J Hydrol Eng., **21** (2016). doi: 10.1061/(asce)he.1943-5584.0001264
19. T. Ouarda, C. Girard, G. Cavadias, B. Bobée, J Hydrol., **254** (2001). doi: 10.1016/s0022-1694(01)00488-7
20. L. Kaufman, P. Rousseeuw, *Finding Groups in Data : An Introduction To Cluster Analysis*, (John Wiley & Sons Inc.,New Jersey, 1990). doi: 10.1002/9780470316801
21. M. Leclerc, T. Ouarda, J Hydrol., **343** (2007). doi: 10.1016/j.jhydrol.2007.06.021
22. L. Xiong, T. Du, C. Xu, S. Guo, C. Jiang, C. Gippel, Water Resour Manag., **29** (2015). doi: 10.1007/s11269-015-1019-6
23. L. Yan, L. Xiong, S. Guo, C. Xu, J. Xia, T. Du, J Hydrol., **551** (2017). doi: 10.1016/j.jhydrol.2017.06.001
24. D. Glatfelter, Techniques for estimating magnitude and frequency of floods on streams in Indiana, *US Geol Surv Water Resour Investig Rep*. (1984). doi: 10.3133/wri844134
25. J. Hosking, J. Wallis, *Regional Frequency Analysis: An approach based on L-moments*, (Cambridge University Press, New York, 1997). doi: 10.1017/cbo9780511529443
26. A. Reynolds, G. Richards, B. de la Iglesia, J Math Model Algorithms., **5** (2006). doi: 10.1007/s10852-005-9022-1
27. E. Nash, V. Sutcliffe, J Hydrol., **10** (1970). doi: 10.1016/0022-1694(70)90255-6
28. R. David, G. Legates, Water Resour Res., **35** (1999). doi: 10.1029/1998WR900018
29. D. Radbruch-Hall, R. Colton, W. Davies, I. Lucchitta, B. Skipp, D. Varnes, *Landslide Overview Map of the Conterminous United States*, (Washington, DC: United States Geological Survey, 1982). doi: 10.3133/pp1183