2012 International Conference on Future Environment and Energy
IPCBEE vol.28(2012) © (2012)IACSIT Press, Singapoore

Assessment of Regional Floods Using Cluster Analysis and Region of Influence (ROI) Method

Isameldin A. Atiem [1], Badreldin G. H. Hassan [1] and Feng Ping[2+]

[1]Nyala University, Nyala, Sudan, e-mail: isam_eldin@hotmail.com
[1]TianjinUniversity, Tianjin, China, e-mail: badr36hassan@yahoo.com
[2] Tianjin University, Tianjin, China, e-mail: fengping@tju.edu.cn

Abstract. In this study, the Region of Influence, ROI, and clustering analysis models are applied to data available for the Nile River. Data analysis was completed for annual extreme flows of a network of 14 stream gauging stations. The methodology consists of: selection of a set of attributes to be included in the distance metric measure among the array of possible advance postulated attribute sets, using a screening process; determination of appropriate weighting values for each selected attribute; performing cluster analysis in attribute-space for 1 to K clusters; computation and plotting of error criterion and F-statistic for the clusters; selecting an appropriate number of clusters; incorporating the ROI approach by determining a threshold value to define a cutoff for the inclusion of stations into ROI for a site; the determination of a weighting function that reflects the relative closeness to the site of each station in a site's ROI, corresponding to options 1, 2, and 3 of the ROI approach; having defined the ROI regions, performing the R-statistic homogeneity test for both clustering and ROI optional regions; partitioning the normalized regional range in the coefficient of variation values for each region; identification of the fixed regions that form the most homogeneous sites in the region; and finally, estimation of their dimensionless growth curves.
The performance of the models is summarized in terms of the variability of the at-site values about the regional curve, root mean squared error (RMSE), and bias. Results indicate that ROI option 1 performs much better than ROI options 2 and 3, and the clustering options.

Key words: PWMs/GEV, Cluster Analysis, ROI, Regionalization, Variability Measure

1. INTRODUCTION

The classical methods in flood frequency analysis are hampered by insufficient gauging networks and insufficient data, especially when the interest is in estimating events of large return periods. At-site flood frequency analysis is the approach where only flood records at the subject site are used. More commonly, it is necessary to carry out a regional analysis in which flood records from a group of similar catchments are used. Regional flood frequency analysis, RFFA, is a probabilistic method which attempts to respond to the need for flood estimation in ungauged basins and to improve the at-site estimate by using the available flood data within a region. Thus, it enables flood quantile estimates for any site in the region to be expressed in terms of flood data recorded at all gauging sites in the same region, including those at the specific site.

RFFA provides a mechanism of exploiting the obvious spatial coherence of hydrologic variables; and, thus, all available relevant information can be incorporated in the flood estimate. Two parameter distributions are not sufficiently flexible to model all plausible flood-like parent distributions, leading to quantile estimates with not excessively large standard errors, but possibly with excessively large bias. On the other hand, three-parameter distributions are sufficiently flexible to be relatively unbiased but accompanied by unacceptably large standard error. These are true even in the case of homogeneous regions and mildly heterogeneous regions (Cunnane, 1988). Kuczera (1982), Lettenmaier & Potter (1985), Lettenmaier et al. (1987), Hosking & Wallis (1986 & 1988) and Burn (1990) have

+ Corresponding author. Tel.: +86 13820121008; fax: +86 22 .27890459
E-mail address: fengping@tju.edu.cn

shown that, in cases of extreme regional heterogeneity, estimates based on RFFA could be preferable instead of at-site estimates.

The starting point of most regional flood frequency procedures is the U.S. Geological Survey index flood method proposed by Dalrymple (1960). Hosking et al. (1985), Wallis & Wood (1985), Hosking & Wallis (1988), Lettenmaier & Potter (1985), Jin & Stedinger (1989) have demonstrated that index flood procedures yield suitably robust and accurate quantile estimates. Burn (1989) explored the problem of delineating a set of stations that can be considered to constitute a homogeneous region, based on relevant basin attributes, using the cluster analysis technique. Burn (1990) implemented the region of influence (ROI) approach. Zrinji & Burn (1996) introduced a hierarchical feature as a refinement to the ROI.

In this study, the generalized extreme value (GEV) distribution is proposed to be used in conjunction with the method of PWMs. Potter (1987) identified that such combination is an efficient basis for combining extreme flow data (Burn,1989 & 1990; Gabriele & Arnell, 1991; Zrinji & Burn, 1996; Lettenmaier et al., 1987; Hosking et al., 1985).

2. METHODOLOGY

2.1. Pooling Regional Cluster Groups

Cluster analysis is used to classify the data in order to capture a diversity of factors, and the K-means algorithm is specified for assigning stations to a cluster or group. The basis of the K-means algorithm is the partitioning of N objects (stations) into K groups, based upon the values of M features, or attributes, of the object. The starting point for the K-means algorithm is an initial centroid for each cluster. The objects are then assigned to the cluster with the nearest centroid in terms of a weighted Euclidean distance measure in M-space. The centroids for each cluster are next recalculated, and objects may be reassigned to different clusters, depending on the distance measure from the object to the new centroid locations (Burn, 1989). This process is repeated until no cluster experiences a change in membership; and the objective function of the clustering algorithm is therefore to minimize the error criterion defined as (Burn, 1989):

$$EC = \sum_{k=1}^{K} \sum_{i \in I_k} \sum_{m=1}^{M} w_m (X_{m,i} - C_{m,k})^2 \tag{1}$$

where EC is the error criterion; wm is the weight applied to attribute m in the distance measure; $X_{m,i}$ is the value of station or object i on attribute m; Cm,k is the centroid coordinate of cluster k for attribute m; K is the number of clusters; Ik is the set of objects in cluster k; and M is the total number of attributes. To determine an appropriate number of clusters, a plot of the EC versus the number of clusters may be examined for points where there are diminishing returns for the decrease in the EC value with an increase in the number of clusters. A further aid is the use of a statistic suggested by Galeatti et al. (1986) defined as:

$$F = \left[\frac{EC(K)}{EC(K+1)} - 1 \right] (N - K + 1) \tag{2}$$

where EC(K) is the error criterion for K clusters; EC(K+1) is the criterion for K+1 clusters; and N is the total number of objects or stations. The value of F is a measure of the reduction in variance from K to K+1 clusters. An empirical rule, a value of F greater than 10, justifies a transition from K to K+1 cluster (Burn, 1989).

2.2. Pooling Regional ROI Groups

The fundamental premise of the ROI approach is that there is no need for distinct boundaries between different regions; rather, in the estimation of at-site extremes, each site should utilize information from all stations that are sufficiently similar to it (Burn, 1990). The ROI approach allows for a potentially unique set of gauging stations to be used in the at-site estimation of extremes for every station in a collection of gauging stations. Burn (1990) argued that, although the identification of homogeneous regions will often lead to an effective and efficient spatial transfer of information, problems and inconsistencies still exist, as in the case of a station which lies on the boundary between two homogeneous regions. Such a station could be regarded as being a partial member in both of the regions that it borders (Acreman & Wiltshire, 1987). Burn (1990) expanded this concept so that there is no need to define boundaries between regions, but rather each site can have its own "region", consisting of those stations that are sufficiently similar to the site of interest.

The starting point of the ROI to regionalization is the selection of a distance metric, defining the closeness of each station to every other station. The similarity among basins is measured by means of a weighted Euclidean distance in

the M-dimensional space, where M represents the number of attributes used to define station similarity. The distance metric is:

$$D_{ij} = [\sum_{m=1}^{M} w_m (X_{m,i} - X_{m,j})^2]^{1/2} \quad (3)$$

where Dij is the weighted distance between station i and station j; wm is the weight applied to attribute m, reflecting the relative importance of the attribute; Xm,i is the value of attribute m for station i; and Xm,j is the value of attribute m for station j.

It is desirable to include attributes with diverse characteristics, such as the ability to combine measures of similarity between the extreme flows data collected at the stations, as well as geographic and physical attributes of the station locations (Burn, 1989). Having defined an appropriate set of attributes, two tasks remain to complete the definition of a station's ROI. The first is to determine a threshold value to define a cutoff for the inclusion of stations into ROI for a site. Any station with a distance value in excess of the threshold will not be included in a site's ROI. The set of stations in a site's ROI are defined as:

$$I_i = \{J : D_{ij} \le \theta_i\} \quad (4)$$

where I_i is the set of stations in the ROI for site i, and θi is the threshold for site i.

The second is a weighting function that reflects the relative closeness (in M-dimensional attribute space) to the site of each station in a site's ROI (Burn, 1990). The weighting function has the form:

$$\eta_{ij} = f(D_{ij}, \Psi) \ldots\ldots\ldots \forall j \in I_i \quad (5)$$

$$\eta_{ij} = 0 \ldots\ldots\ldots \forall j \notin I_i \quad (6)$$

where ηij is the weight for station j in the ROI for site i; f ( ) is a functional relationship defining the weight; and Ψ is a parameter vector for the weighting function.

The weighting function is used in the pooling of information from all stations in the ROI for a site. The characteristics of the ROI approach to regionalization can be captured by formulating numerous options. One possible approach would be to choose a sufficiently large threshold value such that all stations are included in the ROI of every other station. The opposite extreme would result from choosing a very restrictive threshold such that the size of Ii is quite small. Thus, the strategies for formulating ROI options represent different philosophies for combining information for regional flood frequency analysis.

Option 1;

This option has a threshold value defined as:

$$\theta_i = \theta_L, \ldots\ldots\ldots NS_i \ge NST \quad (7)$$
$$\theta_i = \theta_L + (\theta_U - \theta_L)(NST - NS_i)/NST, \ldots\ldots NS_i < NST$$

and a weighting function defined as:

$$\eta_{ij} = 1 - (D_{ij}/TP)^n \quad (8)$$

where θL is a lower threshold value defining a desired proximity for stations to be included in the ROI for site i; NSi is the number of stations in the region of influence for site i with the threshold at θL; NST is the target number of stations for a ROI; θU is an upper threshold value for sites with fewer than NST stations in the ROI; and TP and n are parameters of the weighting function.

Option 2;

This option assumes a constant threshold value given as:

$$\theta_i = \theta_U \quad (9)$$

and a weighting function assumed as:

$$\eta_{ij} = 1 - \left(\frac{D_{ij} - \theta_L}{TN - \theta_L}\right)^n \ldots\ldots\ldots D_{ij} > \theta_L \quad (10)$$
$$\eta_{ij} = 1 \ldots\ldots\ldots\ldots\ldots otherwise$$

where TN and n are parameters of the weighting function with TN defined as:

$$TN = \text{Max}(TL_i, TPP) \quad (11)$$

With TPP a parameter of the weighting function, and TLi given as:

$$TL_i = Max_{\{j\}}(D_{ij}) \quad (12)$$

Option 3;

This option involves all sites in the ROI such that:

$$\theta_i = TL_i \tag{13}$$

and the weighting function is defined as in option 2.

The first option includes a limited number of stations in the ROI for each site, and the resulting stations are then expected to be very similar in the extreme flow response to the site of interest. Options 2 and 3 represent variations on the opposite approach to ROI formulation in that a comparatively large number of stations are included in each ROI (for option 3, all stations). The weighting function is then used to reflect the relative proximity of stations.

With the definition of the station membership for each region of influence and the determination of the weight assigned to each station in the ROI, it is possible to estimate at-site extremes by incorporating information from all stations in the ROI. The methodology for combining information from all of the included stations will necessarily be somewhat specific to the distribution function selected for extreme flows and the parameter estimation technique used (Burn, 1990).

2.3. GEV and Probability Weighted Moments (PWMs)

For operational purposes, at-site sample values of PWMs from a sample Xi, i=1,.., n, are satisfactorily provided by:

$$M_{1r0} = \frac{1}{n}\sum_{i=1}^{n} w_i X_{i,n} \tag{14}$$

$$M_{10s} = \frac{1}{n}\sum_{i=1}^{n} (1-w_i)^s X_{i,n} \tag{15}$$

where Xi:n represents sample values ranked from the smallest to the largest; M100 is the ordinary sample mean; and wi = (1-0.35)/n is a weighted function.

The PWM method of parameter estimation consists of deriving expressions for M1r0 or M10s in terms of parameters of the assumed parent distribution. Cunnane (1988) equated as many of these as there are unknown parameters to sample values calculated from data by the above equations, and he solved the resulting equations for the unknown population parameters. For regional estimation, Wallis (1980) has proposed that at-site values of PWMs be standardized by division by the at-site mean, M100, and the resulting standardized values be averaged across the sites in the region as:

$$\begin{aligned} m_r &= M_{1r0} / M_{100} \\ m_s &= M_{10s} / M_{100} \end{aligned} \tag{16}$$

He then calculated mr and ms for each site and averaged across the M sites by:

$$\bar{m}_r = \sum_{j=1}^{M} m_{r,j}\left(n_j / L\right), \bar{m}_s = \sum_{j=1}^{M} m_{s,j}(n_j / L) \tag{17}$$

while the contribution of each site to the average is weighted in proportion to its record length L. The parameters of the distribution can be estimated using three PWMs obtained from the sample data through the following (Hosking et al., 1985):

$$M_r = \frac{1}{n}\sum_{i=1}^{n} w_i^r X_i \ldots\ldots\ldots r = 0,1,2 \tag{18}$$

where wi = (i-0.35)/n is the plotting position for data point xi and n is the number of flood record at the station. Having calculated the PWMs for each station, the scaled values can then be obtained through:

$$\begin{aligned} t_{1,j} &= M_{1j} / M_{0j} \\ t_{2j} &= M_{2j} / M_{0j} \end{aligned} \tag{19}$$

where j denotes the station number. PWMs for the ROI are then derived from the scaled PWMs of the stations in the ROI through:

$$T_{ki} = \sum_{j \in I_i} t_{k,j} n_j \eta_{ij} / \sum_{j \in I_i} n_j \eta_{ij} \ldots\ldots k = 1,2 \tag{20}$$

where Ii is the set of stations in the ROI for site i, and nj is the number of record for station j. The index i on the regionalized PWM indicates the site for which the weighted PWM is calculated. From the weighted PWM values for each ROI, the parameters of the GEV distribution can be estimated through (Hosking et al., 1985).

2.4. HOMOGENEITY TEST

Wiltshire (1986) explored different measures of regional homogeneity and advocated the use of a distribution-based test which involves the calculation of an R- statistic to measure the variability of the stations included in the region

Wiltshire (1986) demonstrated that R follows the chi-squared distribution with N-1 degrees of freedom. The calculated sample values of R can be compared with tabulated values at a selected probability level to test the hypothesis that the region is homogeneous. Lower values for R justify greater homogeneity. If the resulting regions are not sufficiently homogeneous, the relative weight values in the distance measure may be varied, i.e, to identify weighting combinations that enhance the homogeneity of the regions. To further evaluate the fixed regions, the normalized regional range in the coefficient of variation, CV, values are also added as a convenient measure of regional heterogeneity as (Lettenmaier et al. 1987)

3. APPLICATION

3.1. Data Set and Attribute Selection

In the presented study, the above methodology is applied to a network of 14 stream gauging stations in the Nile River basin. The Nile River is known as a river with extremely high variability. Three major streams, the Blue Nile, the White Nile, and Atbara River, are responsible for the main floods on Nile. The annual river inflow at Aswan varies from year to year; the river rises to about 150 billion m3 during high floods and drops to about 42 billion m^3 in low flood periods. Recorded history of the Nile floods goes back to 3200 B.C, when King Menes of Egypt diverted the Nile by constructing a series of dykes on the river to protect his new capital Memphis, 22 km south of Cairo. More recently, in the flood of 1946, an unforeseen relief to Egypt occurred; however, it caused great damage in the Sudan plains, where the main Nile between Khartoum and Atbara overtopped its banks and flooded large areas. Over the last years, Sudan has encountered a number of high floods on the Nile River and its tributaries, such as the floods of 1988, 1994, 2000, and 2003. Recent floods caused loss of life and substantial damage to the agricultural sector. For example, in the 1988 flood of Sudan, more than 90 people were killed and about 1.5 million people were left homeless.

Figure 1 shows a general layout of the Nile basin and the selected gauging stations numbered from 1 to 14. Table 1 shows some statistical hydrologic and physical characteristics for the selected gauging stations. The selection of attributes is related to the extreme flow response at a particular station. However, reliable estimates of the attribute values should be obtained from the available database, which comprises the annual flow records and limited information describing physical features of the contributing drainage area.

Table 1. Hydrologic and physical characteristics for the study area

Station Name (number)	N	Mean discharges (m^3/s)	Max. discharges (m^3/s)	Min. discharges (m^3/s)	Drainage area (10^3 km^2)	Longitude (^{O}E)	Latitude (^{O}N)
Atbara (1)	77	2795.9	4470	367	123.00	34.00	17.67
Hassanb (2)	86	7612.4	10926	3240	1888.19	33.83	17.49
Tamniat (3)	89	7596.7	11000	3910	1882.63	32.59	15.93
Soba (4)	91	7160.9	10700	3420	324.53	32.50	15.67
Abuhraz (5)	93	166.3	264	45	32.00	33.49	14.47
Hiletidris (6)	94	542.2	1023	99	18.16	33.63	14.00
Sennar (7)	82	6840.6	10900	3691	263.63	33.60	13.60
Wedelis (8)	81	6972.9	11300	3460	259.59	34.07	12.94
Roseires (9)	86	6738.6	11255	3132	250.19	34.40	11.85
Mogran (10)	85	1406.1	2360	924	1573.47	32.40	15.67
Malakal (12)	94	1371.7	2450	1070	1557.07	31.67	9.56
Melut (11)	53	1283.9	2156	645	1561.33	32.14	10.47
Hildlobi (13)	83	785.3	1240	466	225.00	31.60	9.31
Nassir (14)	46	738.8	894	474	207.13	32.62	8.60

3.2. Cluster Analysis

Since the objective function of the clustering procedure is to minimize the error criterion, the process is repeated for clusters K=1, … , 7, where no cluster experiences a change in membership. Prior to performing cluster analysis, the

units of attribute data are standardized in order to eliminate the units from the objects and to relate the spread of the data to the spread of the ensemble of data points for the particular attribute, as recommended by Burn (1989), Burn (1990), Zrinji & Burn (1996), and Bhaskar & O'Connor (1989). Thus, attributes are standardized by dividing attribute values by their means.

For K = 1,…, 7 clusters, error criterions for each cluster group and their corresponding F-statistic values are calculated and plotted in Figure 2. In determining the appropriate number of clusters, the clusters are examined for points where there are diminishing returns for a decrease in the error criterion value with an increase in the number of clusters. Furthermore, it is pursued that the F-statistic value greater than 10 justifies a transition from K to K+1 clusters. Thus, K=3 is suggested to be appropriate, and the number of regions for the 14 stations in the network is set at three.

Fig. 1: Nile Basin and the selected stream gauging stations

3.3. Region of Influence (ROI) Approach

An important part of the procedure for selecting parameter values is the matrix of distances from every station to every other station. The diagonal elements of this matrix are zero; the terms above the diagonal (or the terms below the diagonal) include all observed non-zero distance values. The elements above or below the diagonal are identified by magnitude, resulting in a form analogous to a distribution for weighted distance values between station pairs. Thus, one can determine the median distance, the largest or the smallest distance, or any particular percentile of the distance values. Selected percentiles of the stored distance values are to be used as a guideline for selecting threshold values for the ROI options (Burn, 1990). The selected percentiles act as a guideline because, intuitively, the threshold values should correspond to breakpoints in the array of distance values.

For the weighting function parameters, Burn (1990) considered the modeling approach applied to the individual options. Option 1 involves an ROI containing only those sites that are reasonably similar to the target station. Thus, the weighting function values should be substantially different from zero even at the upper threshold. In contrast, the weighting functions for options 2 and 3 should be given comparatively low weights to stations at the threshold since both of these options entail the inclusion of a large number of stations in the ROI.

In this study, the distance matrix of the M-space attributes is constructed, a plot of Euclidean distance measure is performed, and the breakpoints are defined. Thus, the lower threshold parameter was set to 1.5, which corresponds to 23% percentile; the upper threshold value is set to 3.5, which corresponds to 78% percentile; and the target number of stations (NST) was set at 5 (about one third of the available objects). The parameter values for the weighting function, η_{ij}, were chosen as the following: for option 1, where an ROI, containing sites that are reasonably similar to the target station, η_{ij} should be substantially different from zero even at the upper threshold; thus, n is set to 2.5 and TP is set to 4.1 (91%). For options 2, and 3, the weighting functions should give comparatively low weights to the stations at the threshold since both options entail including a large number of stations in the ROI; thus, n is set to 0.1 and TPP is set to 4.1 (91%).

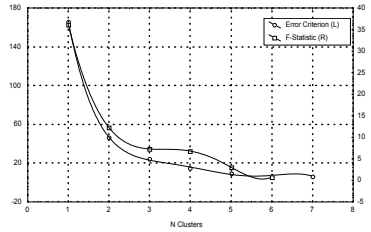

Fig. 2: Plots of the Error criterion and F-statistic

3.4. Testing Regional Groups and Presentation of the Results

Sample R-statistic values are calculated and compared with the tabulated values at 5% level of significance to test the hypothesis that the region is homogeneous. These tests resulted in 5 and 3 ROI regions for options 1 and 2, respectively, and 2 cluster regions for option 1, that passed the homogeneity test at the selected level. To further evaluate the fixed regions, the normalized regional range (R*(CV)) in the coefficient of variation (CV) is used as a measure of regional heterogeneity. Calculation of the normalized regional range for the entire network of 14 stations and the 10 parent regions resulted in 4 ROI regions; and 1 cluster region passed the test of normalized regional range, as shown in Table 2. Fixed regions are re-numbered, region number followed by the letter F and the homogeneity measure rank following the letter F.

Table 3 shows the estimates for dimensionless parameters of the region for the GEV distribution, where the shape parameters were examined. They provide well estimates for the GEV as Hosking et al. (1985) demonstrated ($-0.5 \le g \le 0.5$). It is interesting to note that, in ROI options, the values of the shape parameter, g, remain within the above specified range, as shown in Table 3. Furthermore, the calculated g values for ROI options follow the homogeneity rank, i.e., regions 5F1, 3F2, 2F3, and 2F4, which resulted in shape values of 0.153, 0.176, 0.219, and 0.319, respectively.

Table 2. Partitioning of stations and their assessment

ROI partitioning					
Option No.	ROI No.	Number of sites & the sites included in a region	Sample R	Critical R at 5% level of significance	R*(CV)
1	1	8 (1,3,4,6,10,11,13,14)	11.93	14.07	1.278
	2 F3 *	8 (3,1,4,6,7,8,9,11)	5.68	14.07	0.821
	3 F2 *	10 (4,1,2,3,6,7,8,9,10,11)	7.68	16.92	0.821
	4	11 (6,1,3,4,7,8,9,10,11,13,14)	13.07	18.32	1.254
	5 F1 *	6 (9,7,3,4,6,8)	1.07	11.07	0.672
2	1	12 (1,2,3,4,6,7,8,9,10,11,13,14)	13.14	19.68	1.278
	2 F4 *	10 (7,1,3,4,6,8,9,10,11,13)	10.99	16.92	0.847
	3 *	8 (8,1,3,4,6,7,9,11)	5.68	14.07	0.821
3	1	14 (1,2,…,14)	23.27	22.36	1.231
Partitioning for Cluster Analysis					
Option No.	Cluster number	Number of sites & the sites included in a region	Sample R	Critical R at 5% level of significance	R*(CV)
1	1 *	6 (9,3,4,6,7,8)	1.07	11.07	0.672
	2	6 (1,10,2,11,13,14)	11.11	11.07	1.281
	3 F5 *	2 (5,12)	0.032	3.84	0.04
2	1	14 (1,2,…,14)	23.27	22.36	1.231

* regions which passed the homogeneity tests.

3.5. Tests for Variability and Performance Measures

Sites in each homogeneous region were examined in terms of flood quantiles. A dimensionless flood quantile for return periods (T) of 25, 50, 75, 100, 200, 500, and 1000 year events are calculated, as shown in Table 4 and Figure 3. Variability of the at-site values about the regional curve test is carried out through (Burn, 1989):

$$V_k^T = \frac{1}{NS_k}\sum_{i \in I_k}\left(QS_i^T - QR_k^T\right)^2 \tag{29}$$

where V_k^T is the variability measure for the T-year event for region k; QS_i^T is the at-site estimate of the standardized T-year event for site i; QR_k^T is the regional growth curve estimate for the T-year event for region k; and the summation is over all sites in region.

Performance measures were also analyzed and evaluated in terms of root mean squared error (RMSE) and bias for at-site estimates and the regional estimates for return periods of 25, 50, 75, 100, 200, 500, and 1000 years, as defined through:

$$RMSE_R^T = \left[\frac{1}{NS}\sum_{i=1}^{NS}\frac{1}{NRPS}\sum_{T \in RPS}\left(\frac{QS_i^T - QR_k^T}{QR_k^T}\right)\right]^{1/2} \tag{30}$$

$$BIAS_R^T = \frac{1}{NS}\sum_{i=1}^{NS}\frac{1}{NRPS}\sum_{T \in RPS}\left(\frac{QS_i^T - QR_k^T}{QR_k^T}\right) \tag{31}$$

where $RMSE_R^T$ is the average root mean squared error; $BIAS_R^T$ is the average bias; NS is the number of sites in the region; RPS is the selected return period; and NRPS is the number of return periods selected. The performance of the ROI options is compared to the performance of the traditional clustering method and to the estimates, using all available sites as one region. Table 5 summarizes the results of the two performance measures for each selected region.

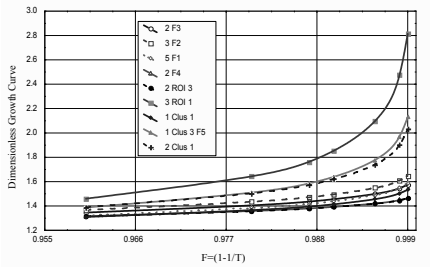

Fig. 3: Regional Dimensionless Growth Curves

4. Discussion of Results and Conclusions

Selected regions are evaluated, based on the R-statistic and the R*(CV) tests, respectively. The normalized regional range for the entire network of 14 stations is 1.231. For the fixed regions, the procedure resulted in values of 0.672, 0.821, 0.821, and 0.847 for regions 5F1, 3F2, 2F3, and 2F4, respectively. For the cluster region 3F5, this measure takes on a value of 0.04. This is due to the similarity between the sites since this region consists of only two sites which may be considered as a single site. In both tests, homogeneity ranks of the regions take the form 5F1, 3F2, 2F3, and 2F4, respectively. One may conclude that ROI approach option 1 forms the most homogenous group, compared to options 2 and 3. For example, the region 2F4 in option 2 is region 3F2 in option 1, excluding site 2 (Hassanb) from the latter region and including site 13 (Hildlobi) in the former region. This may explain the sensitivity of the homogeneity of the region to inclusion or exclusion of a particular site.

Variability about the regional growth curve increases as the return period increases. It is interesting to note that the sum of variability measures in each region follow the homogeneity measures between the regions. Thus, the least homogenous region has a higher sum of variability, and the most homogeneous region has a lower sum. However, the measure should be compared for regions with equal number of members. For example, regions 3F2 and 2F4 have 10 members each; 2F4 is the least homogeneous compared to 3F2, and their variability sums are 0.496, and 0.559, respectively. Moreover, the variability sum in regions of option 1 is always smaller than those of option 2. For example, region 2F3 in option 1 is the same as region 3 in option 2; and their variability sums are 0.299 and 0.313, respectively. In general, the variability measures and the quantile analysis have indicated that at-site quantile estimation of flows regarding selected regions is satisfactory.

Analysis of the performance measures indicates that ROI option 1 performs much better than other ROI options and clustering options. On the other hand, there is a competition between clustering options and ROI options 2 and 3. For example, one may compare regions 2 F3, 5F1, and 1 ROI, which are respectively typical to regions 3, 1 Clu1, and 2 Clu2, as listed in Table 5. The first two regions in the former lie in ROI option1, while the first two regions in the latter lie in ROI option 2 and clustering option 1, respectively. It is clear that ROI option 1 is superior to both ROI option 2 and the clustering option in both performance measures. When the entire available network of 14 sites (1 ROI & 1 Clu2)

is considered, the clustering region performs better than ROI region option 3. This is expected since the nature of options 2 and 3 of the ROI involves information from sites with lesser degrees of similarity reflected by the weights assigned to the included sites. In contrast to ROI option 1, which includes a limited number of sites (stations) with higher degrees of similarity. A comparison of regions including all available sites (1 ROI and 1 Clu2) to other regions indicate the inappropriateness of the former regions; large RMSEs and biases are detected in the former compared to RMSEs and biases of the latter regions.

From the results presented above, one may conclude that the ROI is an efficient regionalization approach and can be easily extended to the case of ungauged sites with some revisions to the attributes selected in the distance measure. In doing this, a regression analysis between observed mean annual floods and corresponding drainage area (A km2), longitude (X_{dis}) and latitude (Y_{dis}) at different sites for regions 5 F1, 2 F3 and 3 F2 is carried out, which respectively results in the following derived equations:

$$M_0 = 48256500 A^{0.69} X_{dis}^{-2.5} Y_{dis}^{-3.42}$$
$$M_0 = 0.0006 A^{0.413} X_{dis}^{1.87} Y_{dis}^{1.48} \quad (32)$$
$$M_0 = 0.0007 A^{0.244} X_{dis}^{2.94} Y_{dis}^{0.75}$$

with regression coefficients of 0.93, 0.62, and 0.48, and explained variances of 86%, 39%, and 23%, respectively. To estimate the flood for an ungauged site, one can easily estimate the mean and multiply it by dimensionless regional quantile estimate for the specified return period, extrapolated from the regional growth curve for the specified region.

Table 3. Regional scaled parameters for GEV

Option	Region	T_1^i	T_2^i	Shape (g) parameter	Scale (α) parameter	Location (ξ) parameter
1 Roi.	5 F1	0.5470	0.3815	0.153	0.153	0.933
	3 F2	0.5548	0.3893	0.176	0.182	0.923
	2 F3	0.5534	0.3874	0.219	0.182	0.928
2 Roi.	2 F4	0.5525	0.3855	0.319	0.189	0.938
3 Roi.	1	0.5541	0.3931	-0.210	0.123	0.897
1 Clu.	3 F5	0.5485	0.3855	-0.087	0.128	0.914
	1	0.5470	0.3810	0.194	0.158	0.935
2 Clu.	1	0.5504	0.3870	-0.041	0.139	0.912

Table 4. Variability Measures and Quantile Values for the Regions

Option	1 ROI			2 ROI		3 ROI	1 Clus.		2 Clus.
Region	2 F3	3 F2	5 F1	2 F4	3	1	1	3 F5	1
T (yr)	V_k^T, Variability measure for the T-year event for region k								
25	0.012	0.014	0.01	0.024	0.011	0.026	0.015	0.00	0.015
50	0.023	0.03	0.013	0.035	0.029	0.064	0.032	0.005	0.031
75	0.031	0.041	0.021	0.047	0.038	0.108	0.042	0.014	0.044
100	0.031	0.049	0.036	0.067	0.042	0.152	0.046	0.026	0.057
200	0.051	0.073	0.05	0.086	0.053	0.321	0.055	0.083	0.101
500	0.056	0.122	0.061	0.133	0.061	0.748	0.070	0.289	0.203
1000	0.095	0.167	0.082	0.167	0.079	1.328	0.088	0.649	0.333
Sum	0.299	0.496	0.273	0.559	0.313	2.747	0.348	1.066	0.784
T (yr)	x_T^i, Site i, T-year flow dimensionless growth curve								
25	1.346	1.367	1.321	1.316	1.314	1.460	1.310	1.387	1.390
50	1.405	1.435	1.384	1.359	1.357	1.643	1.366	1.510	1.504
75	1.436	1.471	1.417	1.380	1.378	1.762	1.395	1.585	1.572
100	1.456	1.495	1.440	1.393	1.392	1.853	1.414	1.639	1.620
200	1.499	1.548	1.490	1.420	1.419	2.096	1.456	1.777	1.740
500	1.546	1.609	1.548	1.448	1.447	2.476	1.504	1.971	1.902
1000	1.576	1.648	1.587	1.464	1.463	2.814	1.534	2.129	2.029

Table 5. Regional Performance Measures

Approach	Option	Region	Site numbers	RMSE	BIAS
ROI	1	5 F1	6 *	0.15678	0.01324
		3 F2	10	0.15973	0.02238
		2 F3	8 *	0.15032	0.02201
	2	2 F4	10	0.20762	0.10436
		3	8 *	0.17552	0.07323
	3	1 ROI	14 *	0.23490	-0.14132
Cluster	1	1 Clu1	6 *	0.16359	0.03386
		3 F5	2	0.19329	0.05315
	2	1 Clu2	14 *	0.17807	-0.06839

* Typical regions (considering site numbers) but in different options or approaches.

5. References

[1] Acreman MC, Wiltshire SE. 1987. Identification of regions for regional flood frequency analysis. *EOS 68(44)*:1262, an abstract.

[2] Bhaskar NR, O'Connor CA. 1989. Comparison of method of residuals and cluster analysis for flood regionalization. *J. of Water Resources Planning and Management 115(6)*: 793-808.

[3] Burn DH. 1989. Cluster analysis as applied to regional flood frequency. *J. of Water Resources Planning and Management 115(5)*: 567-582.

[4] Burn DH. 1990. Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research 26(10)*: 2257-2265.

[5] Cunnane C. 1988. Methods and merits of regional flood frequency analysis. *Journal of Hydrology 100*: 269-290.

[6] Dalrymple T. 1960. Flood frequency methods. U.S. Geological Survey. *Water supply paper 1543A*: 11-51.

[7] Gabriele S, Arnell N. 1991. A Hierarchical approach to regional flood frequency analysis. *Water Resources Research 27(6)*: 1281-1289.

[8] Galeatti, G. et al. 1986. Optimization of a snow network by multivariate statistical analysis. *Hydrol. Sci. Journal 31(1)*: 93-108.

[9] Hosking JR, Wallis JR. 1986. Paleoflood hydrology and flood frequency analysis. *Water Resources Research 22(4)*: 543-550.

[10] Hosking JR, Wallis JR. 1988. The effect of intersite dependence on regional flood frequency analysis. *Water Resources Research 24(4)*: 588-600.

[11] Hosking JR, Wallis JR, Wood EF. 1985. Estimation of the generalized extreme-value distribution by the method of probability weighted moments. *Technometrics 27(3)*: 251-261.

[12] Jin M, Stedinger JR. 1989. Flood frequency analysis with regional and historical information. *Water Resources Research 25(5)*: 925-936.

[13] Kuczera G. 1982. Robust flood frequency models. *Water Resources Research 18(2)*: 315-324.

[14] Lettenmaier DP, Potter KW. 1985. Testing flood frequency estimation methods using a regional flood generation model. *Water Resources Research 21(12)*: 1903-1914.

[15] Lettenmaier DP, Wallis JR, Wood EF. 1987. Effect of regional heterogeneity on flood frequency estimation. *Water Resources Research 23(2)*: 313-323.

[16] Potter KW. 1987. Research on flood frequency analysis: *1983-1986. Rev. Geophys. 25(2)*: 113-118.

[17] Wallis JR. 1980. Risk and uncertainties in the evaluation of flood events for design of hydraulic structures. *Piene e Siccita, Guggino E, Rossi G, Todini E (eds). Fondazione Politecnica del Mediter*., Catania, Italy; 3-36.

[18] Wallis JR, Wood EF. 1985. Relative accuracy of log Pearson III procedures. *J. Hydraul. Div. Am. Soc. Civ. Eng. 111(7)*: 1043-1056.

[19] Wiltshire SE. 1986. Regional flood frequency analysis I: Homogeneity statistics. *Hydrol. Sci. J. 31(3)*: 321-333.

[20] Zrinji Z, Burn DH. 1996. Regional flood frequency with hierarchical region of influence. *Journal of Water Resources Planning and management 122(4)*: 245-252.