Scientia Iranica A (2014) 21(5), 1492–1502

SCIENTIA IRANICA

Sharif University of Technology
Scientia Iranica
Transactions A: Civil Engineering
www.scientiairanica.com

Identification of homogeneous regions and regional frequency analysis for Turkey

M. Firat[a,*], A.C. Koc[b], F. Dikbas[b] and M. Gungor[b]

a. *Department of Civil Engineering, Faculty of Engineering, Inonu University, Malatya, Turkey.*
b. *Department of Civil Engineering, Faculty of Engineering, Pamukkale University, Denizli, Turkey.*

Received 4 August 2012; received in revised form 24 October 2013; accepted 10 December 2013

KEYWORDS
Cluster analysis;
Ward's method;
Annual maximum flow;
Flood frequency analysis;
Hydrologic homogenous region.

Abstract. Ward's hierarchical clustering method is applied to classify the annual floods and identify the hydrologic homogeneous regions in Turkey. For this aim, the annual flood data obtained at the 117 gauging stations having data records of 31 years operated by the general directorate of electrical power resources survey and development administration (EIE) throughout Turkey are considered. Discordancy and regional homogeneity measures are applied to test homogeneity of regions identified by Ward's cluster method. Flood frequency analyses for seven sub-groups defined by Ward's clustering method are carried out using various frequency distributions based on index flood and L-moments approaches. The best fit distributions for all sub-regions are identified based on L-moments goodness of fit statistic. The accuracy of results of quantile estimates are evaluated by using relative RMSE% and relative BIAS% through the use of Monte Carlo simulation.

1. Introduction

Planning and management of water resource projects, such as design of dams, spillways and other water structures, needs accurate estimation of the magnitudes and frequencies of natural extreme events. Regional flood frequency analysis is performed for estimation of flood quantiles in different return periods at basins having missing or short data set. The identification of hydrological homogenous regions is usually the most important and difficult step of the regional analysis. For this reason, for a more accurate regional frequency analysis, the regions should be generated by grouping the stations according to hydrological similarities [1]. Clustering analysis methods such as hierarchical and non-hierarchical clustering algorithms have been widely used to identify the homogenous regions for flood frequency analysis [2-10]. Demirel et al. [11] proposed the use of principal component analysis and K-means methods together in the classification of monthly minimum flows of 23 river basins throughout Turkey. Kahya et al. [12] aimed the spatial classification of river flows in Turkey by using K-means method. Kahya and Demirel [13] used three different clustering algorithms, which are single, and complete linkage, and Ward's methods, to classify low flows in Turkey. It is seen in the literature that the results obtained from the flood frequency analysis carried out by using inhomogeneous regions are better than the results obtained from the frequency analysis made by using one station [14,15]. The index flood L-moments approach for flood frequency analysis has been successfully applied for modeling floods and estimations of flood quantiles [16-19]. Kumar et al. [20] carried out flood frequency analysis based on L-moments approach and selected the Generalized Extreme Value (GEV) distribution as appropriate distribution using goodness of fit measure. Saf [21] determined the regional

*. *Corresponding author. Tel.: +90 4223774742; Fax: +90 4223410046*
E-mail addresses: mahmut.firat@inonu.edu.tr (M. Firat); a_c_koc@pau.edu.tr (A.C. Koc); f_dikbas@pau.edu.tr (F. Dikbas); mgungor@pau.edu.tr (M. Gungor)

probability distributions for the annual maximum flood peaks data recorded at 45 stream-gauging sites in the Kucuk and Buyuk Menderes River Basins using the index flood method. Seckin et al. [22] applied the index flood procedure coupled with the L-moments method to the annual flood peaks data taken at all stream gauging stations in Turkey having at least 15-year-long records. The main purpose of this study is to identify the hydrologically homogenous regions using hierarchical cluster method, called Ward's method, and apply the flood frequency analysis for Turkey. Discordancy and regional homogeneity measures are applied to test homogeneity of regions identified by Ward's cluster method. Then flood frequency analyses for these seven sub-regions are carried out using various frequency distributions based on index flood and L-moments approaches.

2. Method

2.1. Ward's method

Cluster analysis may be defined as the grouping and collecting into a set of the variables with the same properties based on the similarities or differences between the feature vectors in a data set. In this study, Ward's method is applied to identify the homogeneous regions and classify the annual maximum flow. Ward's method is a general hierarchical clustering method proposed by Ward [23] and it is known as the "minimum variance method". In this method, the distance measures between clusters are calculated and the variance analysis approach is used to describe the similarity between clusters. The distance measure is defined as the sum of the squares between two clusters given by [24]:

$$\text{ESS} = \sum_{d=1}^{v}\sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} x_{dij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n_j} x_{dij}\right)^2\right), \qquad (1)$$

where k is the number of clusters, n_j is the number of feature vector at jth cluster, and v is the number of variables.

2.2. Discordancy and regional homogeneity test

The relationship proposed for discordancy measure in the region with N stations defined by cluster analysis [1,14,15] is given by:

$$D_i = \frac{1}{3}N_i(u_i - \bar{u})^T A^{-1}(u_i - \bar{u}), \qquad (2)$$

where D_i is the discordancy measure, N is the number of feature vectors, u_i, ($u_i = [t^{(i)}, t_3^{(i)}, t_4^{(i)}]$) is the vector containing the L-moment ratios of station i, $\bar{u}$ is the regional mean of u_i vector, T is the transposition of a vector or matrix, A is the covariance vector, and $t^{(i)}$, $t_3^{(i)}$ and $t_4^{(i)}$ are L-C_v, L-C_s and L-C_k ratios at station i. It is proposed that a station must be ignored if a site's D_i value calculated for a region with more than 15 stations is higher than 3. Homogeneity of groups defined by cluster analysis is statistically evaluated by using regional homogeneity test based on L-moments ratios proposed by Hosking and Wallis [14]. In homogeneity test, H measures are used for testing the homogeneity of the regions [1,14], given by:

$$H_k = \frac{V_k - \mu_{Vk}}{\sigma_{Vk}}, \qquad k = 1, 2, 3. \qquad (3)$$

In this equation, H_k (H_1, H_2 and H_3) is the measure of regional homogeneity for L-C_v, L-C_s, and L-C_k, respectively, V_k is the variation calculated from the regional data based on regional statistics, σ_{Vk} is the standard deviation of the values obtained from the simulation, and μ_{Vk} is the average of these values. To interpret the H_k values and determine the homogeneity of the regions, the following criteria was proposed by Hosking and Wallis [14]: (i) If $H < 1$ then the region is "acceptably homogeneous", (ii) If $1 \leq H \leq 2$ then the region is "possibly homogeneous", (iii) If $H \geq 2$ then the region is "definitely heterogeneous" [1,14]. The mathematical details of discordancy and homogeneity tests can be obtained from the studies published in the literature [1,14,15,25,26].

2.3. Goodness-of-fit measure

A goodness-of-fit statistic is used for determining the candidate distribution for regions defined by cluster analysis [14,15]. The goodness-of-fit statistic is estimated by simulating a large number of kappa-distributed regions with L-moment ratios and regional averages. The standard deviation of the average L-C_k from simulation, σ_4, and bias, β_4, is measured using Eqs. (4) and (5), respectively. The goodness-of-fit statistic, Z^{DIST}, for each candidate distribution is measured by [14,15]:

$$\sigma_4 = \left[\left(\frac{1}{N_{\text{sim}} - 1}\right)\left\{\sum_{m=1}^{N_{\text{sim}}}(t_4^{(m)} - t_4^R)^2 - N_{\text{sim}}.\beta_4^2\right\}\right]^{1/2}, \qquad (4)$$

$$\beta_4 = \frac{1}{N_{\text{sim}}}.\sum_{m=1}^{N_{\text{sim}}}(t_4^{(m)} - t_4^R), \qquad (5)$$

$$Z^{\text{DIST}} = (\tau_4^{\text{DIST}} - \bar{t}_4 + \beta_4)/\sigma_4, \qquad (6)$$

where N_{sim} is the number of simulated regional data sets generated by a Kappa distribution, and τ_4^{DIST} is the average L-C_k values computed from simulation for a fitted distribution. The fit is considered to be adequate if the statistic is sufficiently close to zero, and a reasonable criterion being $Z^{\text{DIST}} \geq 1.64$. If

more than one candidate distribution is acceptable, the distribution with the lowest Z^{DIST} is considered as the most appropriate distribution for this region [15].

2.4. Index flood method

The index flood method proposed by Dalrymple [27], which is used to determine the magnitude and frequency of flood for basins, located at a hydrological homogeneous region. This method is a fundamental assumption of flood data at different sites in a region follow the same distribution except for a scale or an index factor [27]. Typically the index flood is considered as the mean of the annual flood [15]. Suppose a region defined by cluster analysis with N sites and each site i having data records n_i, observed flood series Q_{ij}, $j = 1, 2, ..., n_i$. The value of estimated flood quantile $Q_i(F)$ for non-exceedance probability F is determined by [15]:

$$Q_i(F) = \mu_i.q(F), \qquad 0 < F < 1, \tag{7}$$

where μ_i is the index flood, and $q(F)$ is the regional quantile of non-exceedance probability F. The average of floods at site i, is computed as $\hat{\mu}_i = \bar{Q}_i = \frac{\sum Q_i}{n}$ and the dimensionless rescaled value of $q(F)$ can be obtained by [15]:

$$q_{ij} = \frac{Q_{ij}}{\hat{\mu}_i}, \qquad j = 1, 2, ..., n_i, \qquad i = 1, 2, ..., N. \tag{8}$$

Hosking and Wallis [15] suggest that the parameters $(\theta_1, \theta_2, ..., \theta_p)$ in index flood method are estimated separately at each site. The regional weighted average of quantile estimates at all sites can be computed using Eq. (9):

$$\hat{\theta}_k^{(R)} = \frac{\sum_{i=1}^{N} n_i.\hat{\theta}_k^{(i)}}{\sum_{i=1}^{N} n_i}. \tag{9}$$

In this equation, $\hat{\theta}_k^{(i)}$ is estimated at site i, N is the number of stations and n_i is the data length at site i. The estimated regional growth curve can be demonstrated as $\hat{q}(F) = q(F; \hat{\theta}_1^R, \hat{\theta}_2^R, ..., \hat{\theta}_p^R)$. Moreover, the quantile estimates at site i are computed by combining the estimates $q(F)$ and $\hat{\mu}_i$ using [14,15]:

$$\hat{Q}_i(F) = \hat{\mu}_i.\hat{q}(F). \tag{10}$$

2.5. Estimation of quantiles and assessment of results

Hosking and Wallis [15] proposed an effective assessment method for defining the properties of complex statistical producers such as regional L-moments and Monte Carlo simulation. The details of this analysis can be given as $\hat{q}^{(m)}$, which is the regional growth curve for non-exceedance probability F, and $Q_i^m(F)$, which is the quantile estimates at site i for non-exceedance probability F, are estimated at mth simulation. Then, the relative error of estimated regional growth curve as an estimator of site growth curve at site i, $q_i(F)$, is computed as $\left\{(\hat{q}^{(m)}(F) - q_i(F))/q_i(F)\right\} q_i(F)$ [15]. Moreover, the relative error of quantile estimates for various non-exceedance probabilities at site i are computed as $\left\{(Q^{(m)}(F) - Q_i(F))/Q_i(F)\right\}$. The relative bias and relative RMSE values for quantile estimates at site i can be calculated as $B_i(F) = \frac{1}{M}\sum_{m=1}^{M} \frac{\hat{Q}_i^m - Q_i(F)}{Q_i(F)}$ and $R_i(F) = \left\{\frac{1}{M}\sum_{m=1}^{M}\left[\frac{\hat{Q}_i^m - Q_i(F)}{Q_i(F)}\right]^2\right\}^{0.5}$, respectively. The regional average relative bias, absolute relative bias and regional average relative RMSE values of quantile estimates are respectively computed according to the equations [15]:

$$B^R(F) = \frac{1}{N}\sum_{i=1}^{N} B_i(F), \tag{11}$$

$$A^R(F) = \frac{1}{N}\sum_{i=1}^{N} |B_i(F)|, \tag{12}$$

$$R^R(F) = \frac{1}{N}\sum_{i=1}^{N} |R_i(F)|. \tag{13}$$

3. Study area and data

In this study, the flow gauging stations operated by general directorate of electrical power resources survey and development administration (EIE) in Turkish basins are used for cluster and flood frequency analysis. The upstream conditions of 257 stations operated by EIE throughout Turkey (if there is a regulation structure or a dam in the upstream), natural flow regime and observation period and other characteristics were investigated in detail. It is stated in literature that the stations to be used in the identification of homogeneous regions and regional frequency analysis should have statistically significant data ($n > 30$ years). Considering this, the stations with more than 31 years were decided to be used. As a result, under all the above evaluations and conditions, a total of 117 flow gauging stations with 31 years observation periods between 1968 and 1998 were selected. The location of the stations used in cluster analysis is shown in Figure 1.

4. Analyses

4.1. Cluster analysis

In this study, the Ward's hierarchical clustering method was applied for identification of homogeneous regions using three data sets having various input variables. The data set 1 includes the variables Q, Q_{dk}, E

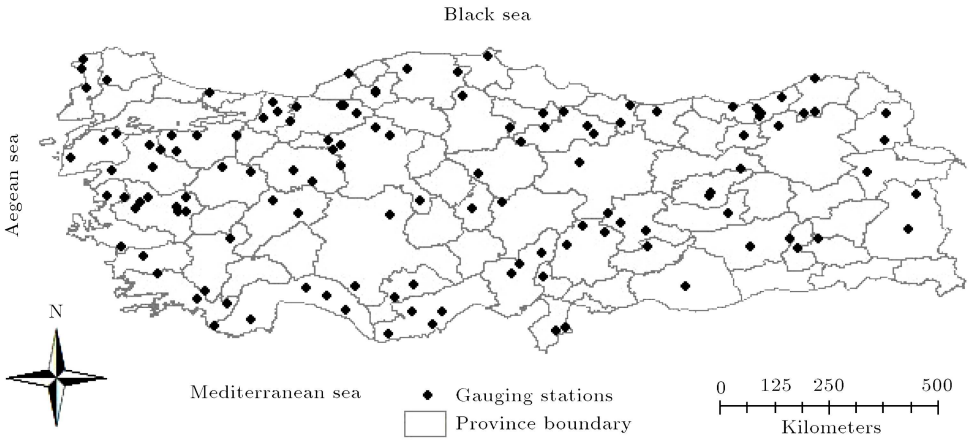

Figure 1. The river flow gauging stations used in cluster analysis.

and B, the data set 2 consists of the variables Q, Q_{dk}, Q_{skew}, E and B, and the data set 3 has the variables Q, Q_{dk}, A, E and B. Q is the annual maximum flow at the station, Q_{dk} and Q_{skew} are the coefficient of variation and skewness coefficient of annual maximum flow at the station, respectively. A is the drainage area and E and B are the latitude and longitude of stations, respectively. The scales of the features used in cluster analysis are very different, and the clustering methods are very sensitive to such scale differences [26]. Equal weight must be assigned to all features, implying equal importance to all the features. Therefore, the features must be transformed so that their ranges are comparable [25,26]. In this study, the data were normalized by using the following transformation function.

$$Q_{yi} = (Q_i - Q_{\min})/(Q_{\max} - Q_{\min}), \tag{14}$$

where Q_i is the annual flow in station i; Q_{yi} is the normalized annual flow in station i; $Q_{\max}$ is the maximum flow in data set, and $Q_{\min}$ is the minimum flow in data set. The optimum number of clusters using Ward's method is determined as 6.

4.2. Discordancy and homogeneity test for regions defined by cluster analysis

The results of regional homogeneity and discordancy tests for the defined regions by cluster analysis are presented in Table 1.

According to the results, the D values for the stations 713, 1323, 2132 and 2232 were found to be higher than the limit value ($D > 3$). When the homogeneity test results for data set 1 are compared, it is seen that the $H1$ value for Regions 1 and 6 are higher than the critical value which is 2. The H values for other regions are lower than the critical value. While the homogeneity test results for Regions 4 and 5 using data set 2 are higher than 2, the H values for other regions is lower than critical value. When the homogeneity test results for data set 3 are compared, the H value for Region 1 is higher than the critical value and for other regions the H values are higher than 2. According to these results, the regions defined by using data set 2 have been selected and used for regional frequency analysis. The distribution of stations in regions defined by cluster analysis is shown in Figure 2.

4.3. Goodness-of-fit measures and choice of frequency distribution

In this study, goodness-of-fit test statistic (Hosking and Wallis, 1997 [15]) was used for determining the suitable distribution for flood frequency analysis in the defined regions (for data set 2) for which homogeneity test was completed. For determining the suitable distributions and other analysis, the software developed by Hosking and Wallis (1993) [14] was used by making new adaptations. The goodness-of-fit test values calculated for regions are presented in Table 2.

In Table 2, the results given for Region 1 show that, while the Z^{DIST} value for the GEV distribution is 0.05, it was obtained to be -1.05 for the GNOR distribution. As indicated before, it is more appropriate to choose the distribution with the lowest absolute value when the goodness-of-fit values for more than one distribution are below critical value. According to this, GEV distribution was chosen as the best fitting distribution for Region 1. The goodness-of-fit test values for Region 2 are 0.31 for GEV distribution, 0.69 for GLOG distribution and -1.62 for GNO and GEV distributions among which the lowest absolute value was chosen as the best fitting distribution. The lowest values of goodness-of-fit test for Regions 3 and 4 were obtained with P3 distribution. In the view of these results P3 distribution was chosen to be best fitting. The values calculated for Regions 5 and 6 indicate that the lowest values are for GEV distribution and GEV distribution was chosen to be best-fitting for these three stations.

Table 1. The results of regional homogeneity test for regions defined by cluster analysis.

Data set	Region	N	H test			Station ID (D measure)
			H1	H2	H3	
Data set 1	1	25	2.863	2.512	2.279	1323(4.41)
	2	12	1.538	0.445	0.258	-
	3	17	-0.606	-0.002	0.067	-
	4	15	1.635	2.009	1.612	713(3.25)
	5	33	1.824	1.206	0.565	2232(4.67)
	6	15	2.964	-0.135	-1.081	2132(3.07)
Data set 2	1	22	1.715	0.448	0.007	-
	2	20	1.495	-0.176	-0.209	713(4.10)
	3	15	1.627	1.873	1.482	-
	4	22	2.419	0.635	0.004	1323(4.47)
	5	15	2.044	0.217	-1.143	2132(3.03)
	6	23	1.471	1.555	0.989	2232(3.44)
Data set 3	1	28	4.288	1.487	0.275	321 (3.12)
	2	10	1.049	0.410	0.268	-
	3	12	1.976	-0.345	-1.537	-
	4	12	2.173	2.143	1.372	-
	5	35	4.054	2.947	1.822	2232 (4.14) 1323 (5.62)
	6	20	6.562	2.839	0.635	2132 (3.61)

Table 2. Goodness-of-fit measures for regions defined by cluster analysis.

Distribution	Z^{DIST}					
	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
GLOG	1.74	0.69	3.97	4.98	1.62	1.57
GEV	**0.05**	**-0.31**	0.97	1.53	**0.34**	**-0.91**
GNO	-1.05	-1.62	0.91	1.36	-0.60	-1.50
P3	-2.97	-3.85	**0.28**	**0.51**	-2.22	-2.73
GPAR	-4.38	-3.41	-5.33	-5.80	-3.08	-6.58

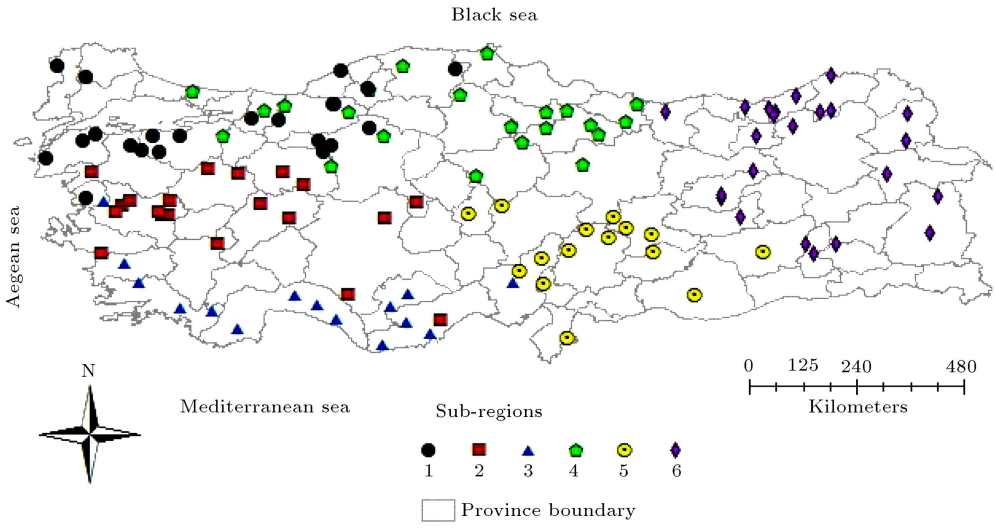

Figure 2. The stations in regions defined by cluster analysis.

4.4. Estimation of quantities with index flood method and evaluation of the results

After determining the hydrologically homogeneous regions and choosing the best-fitting distributions for these regions, regional quantities are estimated. Hosking and Wallis (1997) [15] proposed an effective evaluation method for defining the properties of complex statistical algorithms like L-moments. The estimated regional quantiles for each sub-region are presented in Table 3 with a 90% confidence level.

The investigation of the above results shows that the correctness of the estimations generally decreases

Table 3. The estimated regional quantiles RMSE values at each region.

Region/ distribution	F	T	$\hat{q}(F)$	RSME (%)	Error bounds	
					Lower	Upper
Region 1 /GEV	0.500	2	0.85	0.068	0.801	0.882
	0.800	5	1.357	0.042	1.309	1.371
	0.900	10	1.75	0.065	1.667	1.803
	0.950	20	2.174	0.103	2.014	2.318
	0.980	50	2.803	0.16	2.488	3.134
	0.990	100	3.342	0.206	2.861	3.877
	0.999	1000	5.633	0.384	4.157	7.347
Region 2 /GEV	0.500	2	0.77	0.090	0.700	0.816
	0.800	5	1.382	0.064	1.311	1.401
	0.900	10	1.902	0.066	1.806	1.953
	0.950	20	2.51	0.099	2.315	2.694
	0.980	50	3.492	0.161	3.061	4.009
	0.990	100	4.408	0.212	3.701	5.332
	0.999	1000	9.001	0.400	6.381	12.992
Region 3 /GEV	0.500	2	0.933	0.052	0.899	0.956
	0.800	5	1.372	0.047	1.332	1.4
	0.900	10	1.642	0.064	1.565	1.706
	0.950	20	1.886	0.109	1.766	1.99
	0.980	50	2.186	0.142	2.004	2.343
	0.990	100	2.401	0.183	2.168	2.597
	0.999	1000	3.067	0.401	2.651	3.388
Region 4 /GEV	0.500	2	0.939	0.088	0.909	0.954
	0.800	5	1.354	0.037	1.321	1.374
	0.900	10	1.607	0.071	1.545	1.656
	0.950	20	1.837	0.122	1.738	1.915
	0.980	50	2.118	0.186	1.963	2.234
	0.990	100	2.318	0.229	2.119	2.462
	0.999	1000	2.938	0.361	2.573	3.16
Region 5 /GEV	0.500	2	0.819	0.1	0.75	0.846
	0.800	5	1.391	0.046	1.325	1.4
	0.900	10	1.844	0.078	1.742	1.886
	0.950	20	2.344	0.128	2.145	2.48
	0.980	50	3.103	0.195	2.694	3.448
	0.990	100	3.768	0.247	3.122	4.346
	0.999	1000	6.72	0.44	4.624	8.732
Region 6 /GEV	0.500	2	0.917	0.056	0.889	0.936
	0.800	5	1.301	0.056	1.26	1.312
	0.900	10	1.567	0.08	1.497	1.602
	0.950	20	1.831	0.106	1.719	1.911
	0.980	50	2.188	0.158	1.996	2.355
	0.990	100	2.465	0.201	2.195	2.717
	0.999	1000	3.451	0.356	2.771	4.074

with the increase of recurrence interval. The calculated regional relative RMSE value increases with the increase of recurrence interval and the highest relative RMSE value is obtained when the recurrence period is 1000 years. It can be told that the relative RMSE values calculated for all regions are at an acceptable level. For evaluating the estimation results, relative error of quantile estimate for F non-exceedance probability, relative bias and relative RMSE, regional mean relative bias ($B^R(F)$) and absolute relative bias values of the estimated quantile and regional mean relative RMSE values are calculated. The estimation results of the regional quantiles for each sub-region determined by clustering method are presented in Table 4 with

Table 4. Regional quantiles and regional growth curves at each region.

Region 1														
	Average quantiles							**Growth curves**						
Criteria	**0.50**	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.999**	**0.50**	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.999**
$B^R(F)$	0.015	0.013	0.01	0.006	0.003	0.002	0.019	0.016	0.013	0.009	0.005	0.00 1	0.00	0.016
$A^R(F)$	0.056	0.038	0.057	0.089	0.134	0.171	0.307	0.056	0.038	0.057	0.089	0.1 35	0.171	0.308
RMSE	0.138	0.132	0.145	0.168	0.211	0.251	0.417	0.068	0.042	0.065	0.103	0. 16	0.206	0.384
0.050 PT	0.913	0.905	0.89	0.867	0.835	0.809	0.727	0.963	0.99	0.97	0.938	0 .894	0.862	0.767
0.950 PT	1.123	1.131	1.144	1.166	1.207	1.246	1.425	1.061	1.037	1.05	1.079	1.127	1.168	1.355
Region 2														
	0.50	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.999**	**0.50**	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.9 99**
$B^R(F)$	0.026	0.021	0.014	0.007	0.002	0.006	0.002	0.029	0.0 22	0.014	0.006	0.005	0.011	0.011
$A^R(F)$	0.068	0.059	0.056	0.081	0.127	0.163	0.298	0.07	0.05 9	0.056	0.081	0.127	0.163	0.297
RMSE	0.194	0.19	0.201	0.219	0.258	0.297	0.467	0.09	0.064	0.066	0.099	0.16 1	0.212	0.4
0.050 PT	0.892	0.886	0.865	0.835	0.79	0.754	0.638	0.944	0.987	0.974	0.932	0.871	0.827	0.693
0.950 PT	1.173	1.177	1.193	1.22	1.268	1.315	1.538	1.1	1.054	1.053	1.084	1 .141	1.191	1.411
Region 3														
	0.50	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.9 99**	**0.50**	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.999**
$B^R(F)$	0.008	0.004	0.004	0.005	0.008	0.011	0.023	0.008	0.004	0.004	0.005	0.00 8	0.011	0.024
$A^R(F)$	0.047	0.033	0.055	0.079	0.106	0.124	0.177	0.047	0.033	0.056	0.079	0.10 6	0.124	0.177
RMSE	0.105	0.093	0.108	0.126	0.15	0.167	0.219	0.052	0.037	0.064	0.091	0.1 22	0.143	0.201
0.050 PT	0.914	0.92	0.917	0.913	0.908	0.903	0.891	0.976	0.98	0.962	0.948	0.933	0.925	0.905
0.950 PT	1.104	1.09	1.092	1.102	1.117	1.129	1.171	1.038	1.03	1.049	1.068	1.091	1.108	1.157
Region 4														
	0.50	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.9 99**	**0.50**	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.999**
$B^R(F)$	0.009	0.005	0.004	0.006	0.01	0.013	0.029	0.01	0.005	0.005	0.006	0.01	0 .014	0.03
$A^R(F)$	0.044	0.034	0.066	0.094	0.126	0.147	0.204	0.044	0.033	0.065	0.093	0.12 6	0.147	0.204
RMSE	0.100	0.092	0.112	0.135	0.163	0.182	0.238	0.048	0.037	0.071	0.102	0. 136	0.159	0.221
0.050 PT	0.934	0.935	0.932	0.929	0.924	0.922	0.916	0.984	0.985	0.971	0.95 9	0.948	0.942	0.93
0.950 PT	1.088	1.075	1.078	1.086	1.1	1.112	1.151	1.033	1.025	1.04	1.057	1 .078	1.094	1.142
Region 5														
	0.50	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.9 99**	**0.50**	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.999**
$B^R(F)$	0.035	0.022	0.019	0.017	0.018	0.022	0.061	0.035	0.022	0.018	0.015	0.01 6	0.019	0.055
$A^R(F)$	0.085	0.041	0.072	0.113	0.166	0.205	0.35	0.086	0.04	0.072	0.114	0.167	0.207	0.352
RMSE	0.189	0.161	0.179	0.209	0.26	0.305	0.484	0.1	0.046	0.078	0.128	0.195	0.24 7	0.44
0.050 PT	0.914	0.902	0.884	0.863	0.832	0.806	0.727	0.968	0.994	0.978	0.945	0.9	0.867	0.77
0.950 PT	1.165	1.158	1.168	1.194	1.247	1.299	1.54	1.093	1.05	1.058	1.093	1.152	1.207	1.453
Region 6														
	0.50	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.9 99**	**0.50**	**0.80**	**0.90**	**0.95**	**0.98**	**0.99**	**0.999**
$B^R(F)$	0.006	0.011	0.011	0.01	0.008	0.008	0.024	0.006	0.011	0.012	0.01	0.008	0.008	0.024
$A^R(F)$	0.041	0.043	0.053	0.076	0.124	0.163	0.29	0.041	0.043	0.053	0.076	0.125	0.163	0.29
RMSE	0.1	0.104	0.114	0.131	0.168	0.204	0.341	0.046	0.046	0.06	0.086	0.138	0.181	0.326
0.050 PT	0.939	0.94	0.932	0.919	0.897	0.88	0.827	0.979	0.991	0.978	0.958	0.929	0.907	0.847
0.950 PT	1.076	1.086	1.095	1.109	1.133	1.157	1.272	1.031	1.032	1.047	1.066	1.096	1.1 23	1.246

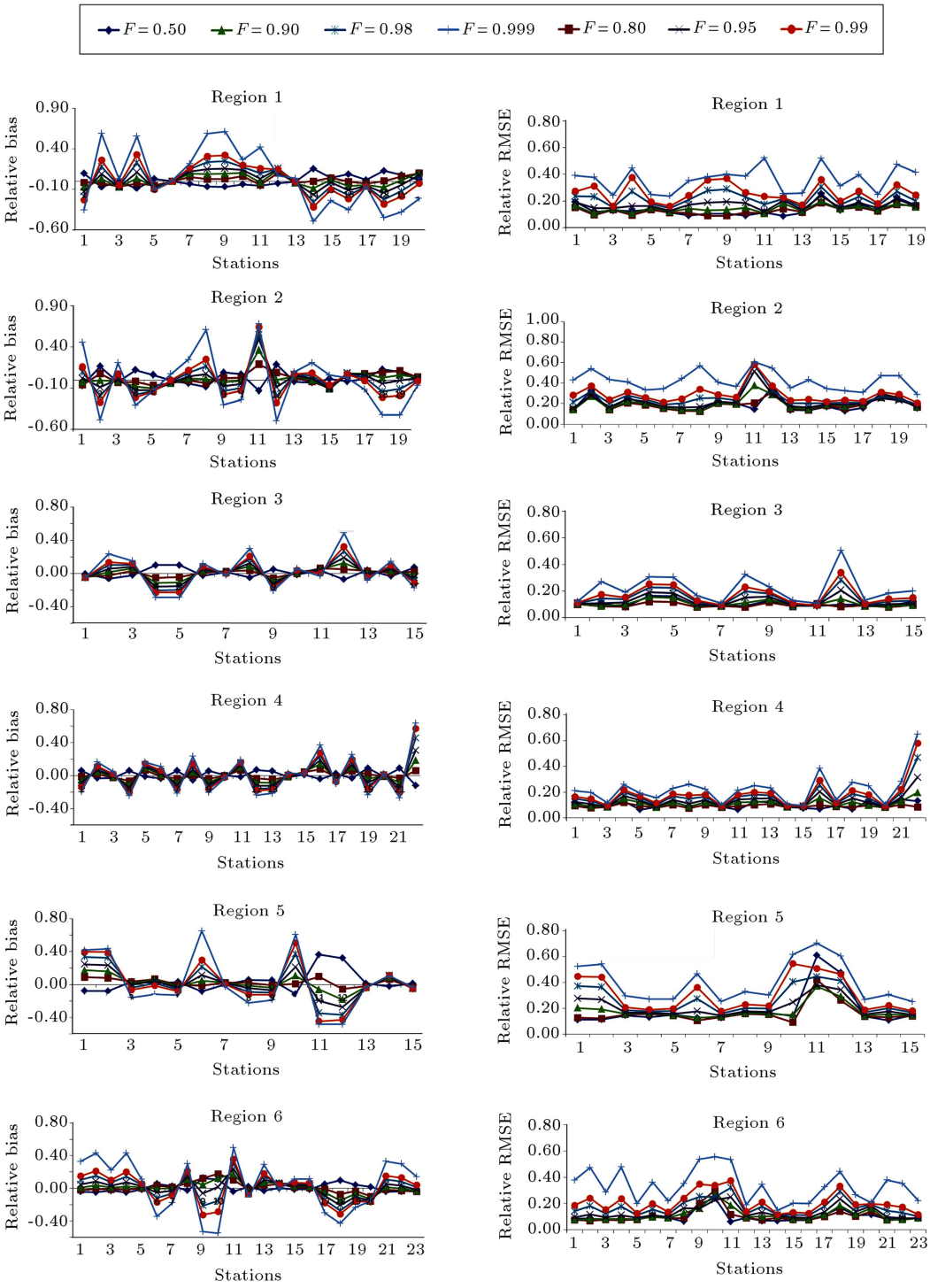

Figure 3. Relative bias and relative RMSE simulation results for the stream gauging stations at each region.

a 90% significance level for various non-exceedance probabilities.

The regional mean values of the criteria calculated for evaluating the estimation results are given in detail in Table 4. Here, criteria for the stations located in each cluster are also calculated. Figure 3 shows the variation of the relative bias and relative RMSE values calculated for the stations located in each region.

In Figure 3, the graphs show that the highest relative RMSE values are generally obtained when the recurrence intervals are high. According to the results obtained by the graphs, the highest relative RMSE value is calculated for the stations in Region 1. The lowest relative RMSE value is obtained for Regions 3 and 7 where the best result in the regional homogeneity test, namely the lowest $H1$ value, was

Table 5. Regional parameters of distributions and growth curve values for each region.

Region	Distribution	Distribution parameters			Growth curves						
	type	ξ	α	K	0.50	0.80	0.90	0.95	0.98	0.99	0.999
1	GEV	0.705	0.384	-0.164	0.85	1.357	1.75	2.174	2.803	3.342	5.633
2	GEV	0.609	0.418	-0.268	0.77	1.382	1.902	2.51	3.492	4.408	9.001
3	P3	1.000	0.48	0.846	0.933	1.372	1.642	1.886	2.186	2.401	3.067
4	P3	1.000	0.463	0.911	0.939	1.354	1.607	1.837	2.118	2.318	2.938
5	GEV	0.676	0.408	-0.181	0.819	1.391	1.844	2.344	3.103	3.768	6.72
6	GEV	0.793	0.33	-0.05	0.917	1.301	1.567	1.831	2.188	2.465	3.451

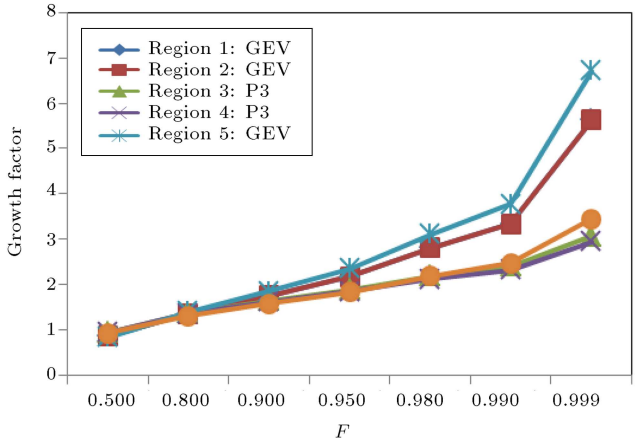

Figure 4. Regional parameters of growth curve values for each region.

obtained. Table 5 shows the regional parameters of distributions at the 90% confidence level and the growth curve values for the selected probability distributions for each sub-region. The regional parameters of growth curve values for each region is shown in Figure 4.

Flood discharges are calculated using distribution function parameters into the probability distribution function with the parameters given in Table 5. According to Table 5, GEV distribution is the most suitable distribution for Regions 1, 2, 5 and 6. The GEV distribution is given by Hosking and Wallis, 1993 [14] and Seckin et al., 2011 [22].

$$Q_T = \left(\xi + (\frac{\alpha}{K}).[1 - (\mathrm{Ln}F)^K]\right).\bar{Q}. \quad (15)$$

The parameters of the GEV distribution can be calculated by:

$$Q_T = (-2.341 + 3.046.(-\mathrm{Ln}F)^{-0.164}).\bar{Q}$$

for Region 1, (16)

$$Q_T = (-1.559 + 2.168.(-\mathrm{Ln}F)^{-0.268}).\bar{Q}$$

for Region 2, (17)

$$Q_T = (-2.254 + 2.293.(-\mathrm{Ln}F)^{-0.181}).\bar{Q}$$

for Region 5, (18)

$$Q_T = (-6.60 + 7.393.(-\mathrm{Ln}F)^{-0.0505}).\bar{Q}$$

for Region 6. (19)

5. Conclusions

The Ward's method was applied for classification of annual maximum flows and identification of hydrologic homogeneous regions. For this aim, the annual maximum river flow data obtained from the 117 stream gauging stations throughout the Turkey were used. Optimum number of clusters for the classification of annual maximum flow was determined to be 6. The regional homogeneity of the clusters was identified by cluster analysis and tested forthcoming regional studies and regional frequency analysis. The discordancy values for stations 713, 1323, 2132 and 2232 were found to be higher than the limit value. On the other side, regional homogeneity test results were evaluated, and it was found that H values calculated for Regions 1, 2, 3, and 6 defined with data set 2 were lower than the limit value of 2. However the H values for Regions 4 and 5 is higher than the critical value. The best fitting distributions for the 6 regions were determined by using goodness-of-fit test statistic. As a result, the best fitting distribution for Regions 3 and 4 was found to be P3 and GEV for Regions 1, 2, 5 and 6. For evaluating the estimation results, relative error of quantile estimate for F non-exceedance probability, relative bias and relative RMSE, regional mean relative bias and absolute relative bias and regional mean relative RMSE values are calculated. The highest relative RMSE values are generally obtained when the recurrence intervals are high, and the correctness of the quantile estimates decrease when the recurrence interval increases. The calculated regional relative RMSE value increases with the increase of recurrence interval, and the highest relative RMSE value is obtained when the recurrence period is 1000. The method produces

successful results in the estimation of flood magnitude in the homogeneous regions determined with the cluster analysis.

Acknowledgment

This research was supported by TUBITAK (Turkish National Science Foundation) under the project number 107Y318. The authors are grateful for editors and anonymous reviewers for their helpful and constructive comments on an earlier draft of this paper.

References

1. Dikbas, F., Firat, M., Koc, A.C. and Gungor, M. "Classification of precipitation series using fuzzy cluster method", *International Journal of Climatology.*, **32**(10), pp. 1596-1603 (2012).
2. Burn, D.H. "Cluster analysis as applied to regional flood frequency", *Journal of Water Resources Planning and Management*, **115**, pp. 567-582 (1989).
3. Burn, D.H. "Catchment similarity for regional flood frequency analysis using seasonality measures", *Journal of Hydrology*, **202**, pp. 212-230 (1997).
4. Guttman, N.B. "The use of L-moments in the determination of regional precipitation climates", *Journal of Climate*, **6**, pp. 2309-2325 (1993).
5. Burn, D.H. and Goel, N.K. "The formation of groups for regional flood frequency analysis", *Hydrological Sciences Journal*, **45**(1), pp. 97-112 (2000).
6. Mosley, M.P. "Delimitation of New Zealand hydrologic regions", *Journal of Hydrology*, **49**, pp. 173-192 (1981)
7. Soltani, S. and Modarres, R. "Classification of spatio -temporal pattern of rainfall in Iran using a hierarchical and divisive cluster analysis", *Journal of Spatial Hydrology*, **6**(2), pp. 1-12 (2006).
8. Lecce, S.A. "Spatial variations in the timing of annual floods in the southeastern United States", *Journal of Hydrology*, **235**, pp. 151-169 (2000).
9. Demirel, M.C. "Cluster analysis of streamflow data over Turkey", Master of Science Thesis İstanbul Technical University, 119 p (2004).
10. Unal, Y., Kındap, T. and Karaca, M. "Redefining the climate zones of Turkey using cluster analysis", *Int. J. Climatol.*, **23**, pp. 1045-1055 (2003).
11. Demirel, M.C., Mariano, A.J. and Kahya, E. "Performing k-means analysis to drought principal components of Turkish rivers", *Hydrology Days*, pp. 145-151 (2007).
12. Kahya, E., Demirel, M.C. and Piechota, T.C. "Spatial grouping of annual streamflow patterns in Turkey", *Hydrology Days*, pp. 169-176 (2007).
13. Kahya, E. and Demirel, M.C. "A comparison of low-flow clustering methods: streamflow grouping", *Journal of Engineering and Applied Sciences*, **2**(3), pp. 524-530 (2007).
14. Hosking, J.R.M. and Wallis, J.R. "Some statistics useful in regional frequency analysis", *Water Resources Research*, **29**(2), pp. 271-281 (1993).
15. Hosking, J.R.M. and Wallis, J.R., *Regional Frequency Analysis: An Approach Based on L-moments*, Cambridge University Press, Cambridge, UK (1997).
16. Lim, Y.H. and Lye, L.M. "Regional flood estimation for ungauged basins in Sarawak, Malaysia", *Hydrol Science*, **48**(1), pp. 79-94 (2003)
17. Vogel, R.M., Thomas, W.O. and McMahon, T. "A flood-flow frequency model selection in southwestern united states", *Journal of Water Resour Planning Management*, **119**(3), pp. 353-366 (1993).
18. Parida, B.P., Kachroo, R.K. and Shrestha, D.B. "Regional flood frequency analysis of Mahi-Sabarmati basin (subzone 3-a) using index flood procedure with L-moments", *Water Resources Management*, **12**, pp. 1-12 (1998).
19. Saf, B., Dikbas, F. and Yasar, M. "Determination of regional frequency distributions of floods in west Mediterranean river basins in Turkey", *Fresenius Environ Bull.*, **16**(10), pp. 1300-1308 (2007).
20. Kumar, R., Chatterjee, C., Kumar, S., Lohani, A.K. and Singh, R.D. "Development of regional flood frequency relationships using L-moments for middle ganga plains subzone 1(f) of India", *Water Resources Management*, **17**(4), pp. 243-257 (2003).
21. Saf, B. "Regional flood frequency analysis using L-moments for the west Mediterranean region of Turkey", *Water Resources Management*, **23**, pp. 531-551 (2009).
22. Seckin, N., Haktanir, T. and Yurtal, R. "Flood frequency analysis of Turkey using L-moments method", *Hydrological Processes.*, **25**(22), pp. 3499-3505 (2011).
23. Ward, J.H. "Hierarchical groupings to optimize an objective function", *Journal of the American Statistical Association*, **58**, pp. 236-244 (1963).
24. Munoz-Diaz, D. and Rodrigo, F.S. "Spatio-temporal patterns of seasonal rainfall in Spain (1912-2000) using cluster and principal component analysis", *Comparison, Annales Geophysicae*, **22**, pp. 1435-1448 (2004).
25. Lim, Y.H. and Voeller, D.L. "Regional flood estimations in red river using L-moment-based index-flood and bulletin 17B procedures", *Journal of Hydrologic Engineering*, **14**(9), pp. 1002-1016 (2009).
26. Cannarozzo, M., Noto, L.V., Viola, F. and La Loggia, G. "Annual runoff regional frequency analysis in sicily", *Physics and Chemistry of the Earth*, **34**, pp. 679-687 (2009).
27. Dalrymple, T. "Flood frequency analyses", *Water*

Supply Paper 1543-A, U.S. Geological Survey, Reston, Va (1960).

Biographies

Mahmut Firat received the BS degree in Civil Engineering from Pamukkale University, Turkey, in 2000 and MSc and PhD degrees in Civil Engineering from Pamukkale University in Denizli, Turkey, in 2002 and 2007, respectively. He is currently an Associate Professor at Civil Engineering Department of Inonu University. His research interests lie in the general area of hydrology, flood analysis, non-revenue water and data analysis.

Abdullah Cem Koc received BS, MS, and PhD degrees in Civil Engineering (Hydraulics, Hydrology and Water Resources Engineering) from the Dokuz Eylul University, Izmir, Turkey, in 1992, 1995 and 2001, respectively. He is currently an Associate Professor at Civil Engineering Department of Pamukkale University, Denizli, Turkey. His research interests include hydraulic transients, water resources management, and hydraulic transients.

Fatih Dikbas received the BS degree in Civil Engineering from Istanbul Technical University, Turkey, in 1994 and MSc and PhD degrees in Civil Engineering from Pamukkale University in Denizli, Turkey, in 1996 and 2002, respectively. He is currently an assistant professor of Civil Engineering Department at Pamukkale University. His research interests lie in the general area of hydraulics, hydrology, data analysis and programming.

Mahmud Gungor received the BS degree in Civil Engineering from Akdeniz University, Turkey, in 1985, MSc in Civil Engineering from Selcuk University and PhD degrees in Civil Engineering from Yildiz Technical University in Turkey. He is currently an Associate Professor at Civil Engineering Department of Pamukkale University, Denizli, Turkey. His research interests lie in the general area of hydraulics, hydrology, and bridge scouring.