An Investigation to Identify Homogenous Regions: A Case Study for New South Wales, Australia

Ayesha S Rahman[1] and Ataur Rahman[2]

[1]PhD Student, Western Sydney University, NSW, Australia
[2]Associate Professor, Western Sydney University, NSW, Australia
Corresponding author's E-mail: ayesha.rahman@westernsydney.edu.au

Abstract

This study focuses on identification of hydrologically homogeneous regions in New South Wales (NSW) from 88 selected sites for regional flood frequency analysis. The heterogeneity measure proposed by Hosking and Wallis (1993) has been adopted for this study amongst the various available techniques to form homogenous region. This method applies L moments to check for homogeneity. L moments for all the 88 sites have been computed and an investigation for discordant sites has been carried out. This has been done in two steps: (i) considering all 88 sites as a single group and (ii) dividing the 88 sites into two groups based on their drainage divisions. The effect of the discordant sites has been examined both individually and as a group. The overall results show that the effect of the discordant sites on heterogeneity measure is negligible and no homogeneous regions can be established in NSW, which is similar to the findings of previous studies.

***Keywords*: ** Homogeneous region, heterogeneity measure, L moments, discordant sites, drainage divisions.

1. INTRODUCTION

Flood is one of the worst natural disasters which can result in millions of dollars of financial damage and loss of life. Design of different types of hydraulic structures, flood plain zoning and economic estimation of flood protection projects and similar tasks requires information of flood magnitude and frequency, which is known as design flood. Resources can be wasted due to the results of under- or over-design of hydraulic structures due to inaccurate design flood estimates. It is hence essential to develop the appropriate models to estimate design floods. Flood frequency analysis (FFA) is commonly used to estimate design floods; however; problem arises in the form of ungauged sites where little or no data is available to employ FFA. Regional flood frequency estimation (RFFE) is usually applied in such cases of ungauged sites with the goal of transferring information from gauged sites to the ungauged target site within a homogeneous region. Therefore, the first priority in RFFE is to identify homogeneous regions and the assessment of the credibility of the obtained regions.

A group of sites may be defined as homogeneous group if their standardised flood frequency curves are similar, within a certain margin of sampling variability (Ribeiro-Correa et al., 1995). Several researchers have proposed homogeneity tests in RFFE including Dalrymple (1960), Wiltshire (1986a, 1986b), Chowdhury et al. (1991), Lu and Stedinger (1992), Fill and Stedinger (1995), and Hosking and Wallis (1993, 1997). The method proposed by Dalrymple (1960) is based on the sampling distribution of the standardised 10 year annual maximum flow, assuming an extreme value 1 (EV1) distribution whereas Wiltshire (1986a, b) presented a test based on the sampling distribution of coefficient of variation of flood data (C_v) and also a distribution-based test. He introduced an F statistic, which is the ratio of between group variance of C_vs and within group variances of C_vs, to judge the degree of homogeneity. However, the problem with these distribution-specific tests is that, when the hypothesis of homogeneity is rejected, it remains doubtful whether the region is heterogeneous, or whether it is homogeneous but has some other parent distribution (Hosking and Wallis, 1993). Nowadays L moments based statistics proposed by Hosking and Wallis (1993, 1997) is very popular in RFFE to form homogeneous regions. L moments are analogous to conventional moments with measures of location (mean), scale (standard deviation), and shape (skewness and kurtosis) and do not involve squaring or cubing the observations, as is done for the conventional method of moments estimators; they are generally more robust and less sensitive to outliers.

Numerous studies have been carried out using L moments based homogeneity analysis. For example, Saf (2009) tested for homogeneity in Turkey based on simulation with a four-parameter Kappa distribution and L-moments based heterogeneity measure. Three subregions have been defined, as the Antalya subregion, the Lower West Mediterranean subregion, and the Upper West Mediterranean subregion. Parida et al. (1998) applied L moments based heterogeneity measure on the 12 gauged sites in India and found one homogeneous region. Application of this same method on 50 stream flow gauging sites in Sicily resulted into 5 homogeneous regions (Noto and La Loggia, 2009). Castellarin et al. (2001) investigated for homogeneous regions by applying the L moments based heterogeneity measure with other methods. Bates et al. (1998), Rahman et al. (1999), Mecivski et al. (2015) attempted to form homogeneous regions in south-east Australia (which included stations from NSW and Victoria) based on Hosking and Wallis (1993) method, but they could not establish any homogeneous regions.

In order to define homogeneous regions to use in RFFE, a trade off exists, whether including additional catchments to make the region larger and maintaining a high degree of similarity with a smaller region. If more sites are added to a region, more knowledge about flood characteristics is available; however, if the added sites are hydrologically dissimilar, the additional information does not result in more precise quantile estimation. With the latest flood data in NSW from 88 sites, this study investigates identification of homogeneous regions using Hosking and Wallis (1993) criteria.

2. STUDY AREA AND DATA SELECTION

For this study 88 catchments from NSW have been selected. The locations of these catchments are presented in Figure 1. These sites are mostly unregulated and have not been affected by major land use changes. All the selected catchments are small to medium in size and the catchment area for the selected sites varies from 8 to 1010 km^2 with a mean of 352 km^2 and median of 260 km^2. Record length ranges from 25 to 82 years, with a mean of 41.5 years and median of 37 years. The physiography of the area varies from mountainous region to coastal area with the mean annual rainfall (MAR) ranging from 625 to 1955 mm/y (mean: 1000 mm/y and median: 910 mm/y).

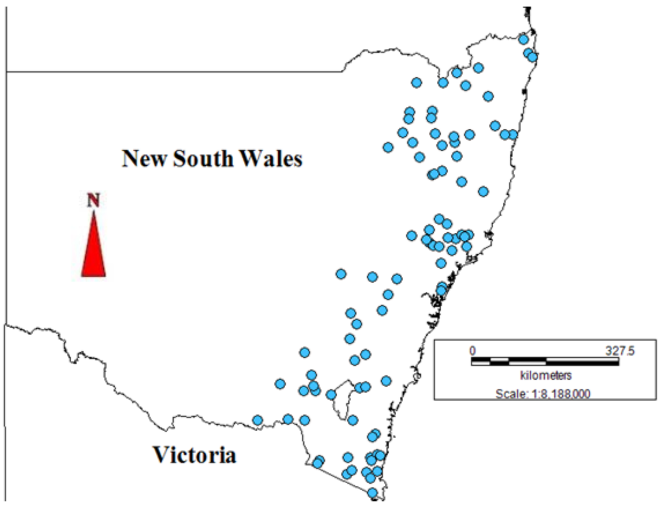

Figure 1. Locations of the selected 88 sites in NSW, Australia

3. METHODOLOGY

For this study, L moments based heterogeneity measure proposed by Hosking and Wallis (1993, 1997) has been selected to identify homogeneous regions from the selected 88 sites. L moments of the annual maximum (AM) flow data are computed for all the selected stations and heterogeneity measure has been applied. The heterogeneity measure looks for discordant sites among the data set. The heterogeneity measure has been applied in two steps: (i) considering all 88 sites in one group and (ii) dividing these selected sites based on their drainage divisions (Drainage Divisions 2 and 4). Drainage Division 2 represents catchments in NSW which are located in the east of the Great Dividing Range, and Drainage Division 4 represents catchments located in the west of the Great Dividing Range. The impact of the discordant sites has been examined both individually and as a group on the data and heterogeneity of the sites has been reassessed.

3.1 L moments computation

Probability-weighted moments (PWMs), introduced by Greenwood et al. (1979) are defined as:

$$\beta_r = E[x\{F(x)^r] \tag{1}$$

which can be rewritten as:

$$\beta_r = \int_0^1 x(F)F^r dF \tag{2}$$

where *F* = *F(x)* is the cumulative distribution function (CDF) of *x*, *x(F)* is the inverse CDF of *x* evaluated at the probability *F*, and *r* = 0, 1, 2, . . . , *s* is a nonnegative integer. When $r = 0$, β_0 is equal to the mean of the distribution $\mu = E[x]$ L Moments are linear combinations of PWMs. The *r*th L moment λ_r is related to the *r*th PWM through:

$$\lambda_{r+1} = \sum_{k=0}^{r} \beta_k (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} \tag{3}$$

For example, the first four L-moments are related to PWMs using

$$\lambda_1 = \beta_0 \tag{4}$$

$$\lambda_2 = 2\beta_1 - \beta_0 \tag{5}$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{6}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \tag{7}$$

Hosking (1990) defined the L-moment ratios as follows:

$$L - C_v = \tau_2 = {}^{\lambda_2}/_{\lambda_1} \tag{8}$$

$$L - skew = \tau_3 = {}^{\lambda_3}/_{\lambda_2} \tag{9}$$

$$L - kutosis = \tau_4 = {}^{\lambda_4}/_{\lambda_2} \tag{10}$$

3.2 Discordancy measure

Let *i* be a site from the proposed region that has *N* number of sites. For site *i*, the vector containing the L moment ratios τ_2, τ_3 and τ_4 are represented as:

$$u_i = [\ \tau_2^{(i)} \quad \tau_3^{(i)} \quad \tau_4^{(i)}] \tag{11}$$

And the unweighted group average $\bar{u}$ is calculated as following:

$$\bar{u} = \frac{\sum_{i=1}^{N} u_i}{N} \tag{12}$$

Then the sample covariance matrix can be found using the following:

$$S = \frac{\sum_{i=1}^{N}(u_i - \bar{u})(u_i - \bar{u})^T}{N-1} \tag{13}$$

A site is considered to be a discordant if it is located far away in the *L* moments space as compared to the majority of the sites in the proposed group. The discordancy measure for site *i* can be measured by the following:

$$D_i = \frac{1}{3}(u_i - \bar{u})^T S^{-1}(u_i - \bar{u}) \tag{14}$$

To identify a site to be discordant, Hosking and Wallis (1993) suggest $D_i \geq 3$. However, they recommend that the data for sites with the largest D_i values should still be examined for possible errors.

3.3 Heterogeneity measure

Let site *i* have record length n_i and sample L moment ratios are $\tau_2{}^{(i)}$, $\tau_3{}^{(i)}$ and $\tau_4{}^{(i)}$. If the regional average *L-C$_v$*, *L-skewness* and *L-kurtosis* weighted by the sites' record length are represented as $\tau_2{}^R$, $\tau_3{}^R$ and $\tau_4{}^R$ where:

$$\tau_R = \frac{\sum_{i=1}^{N} n_i \tau_2{}^{(i)}}{\sum_{i=1}^{N} n_i} \tag{15}$$

The weighted standard deviation can be found as following:

$$V = \{\frac{\sum_{i=1}^{N} n_i(\tau_2{}^{(i)} - \tau_2{}^R)}{\sum_{i=1}^{N} n_i}\}^{1/2} \tag{16}$$

A kappa distribution is fitted to the regional average L moments and a large number of simulations are run to generate N_{sim} of realisations of a region with *N* sites each having the same kappa distribution as their frequency distribution. These simulated regions are homogeneous and have no cross-correlation or serial correlation; they have the same record length as their real counterparts. For each simulated region *V* is calculated. Heterogeneity measure or the *H* statistics can be computed using the following equation

$$H = \frac{(V - \mu_v)}{\sigma_v} \tag{17}$$

Where μ_v and σ_v are the mean and standard deviation of N_{sim} values of V respectively. As suggested by Hosking and Wallis (1993), a region is declared as heterogeneous if H has a large value. Values of H less than 1 can be labelled as 'acceptably homogeneous', values lying between +1 to +2 can mean 'possibly heterogeneous' and $H \geq 2$ indicates a 'definitely heterogeneous' region.

4. RESULTS

4.1 Heterogeneity analysis for the whole data set

Figure 2 shows the discordancy measures D_i for each of the 88 sites computed by equation 14. Amongst these 88 sites four sites (203002, 222016, 401015 and 419029) are found to have D_i values greater than 3. The values of D_i for sites 203002, 401015 and 419029 are 3.78, 3.48 and 3.86, respectively; the value of D_i for site 222016 is found to be the highest i.e. 6.05. Although according to the suggested criteria, these sites could be discordant, further investigations are carried out to be certain about their discordancy status.

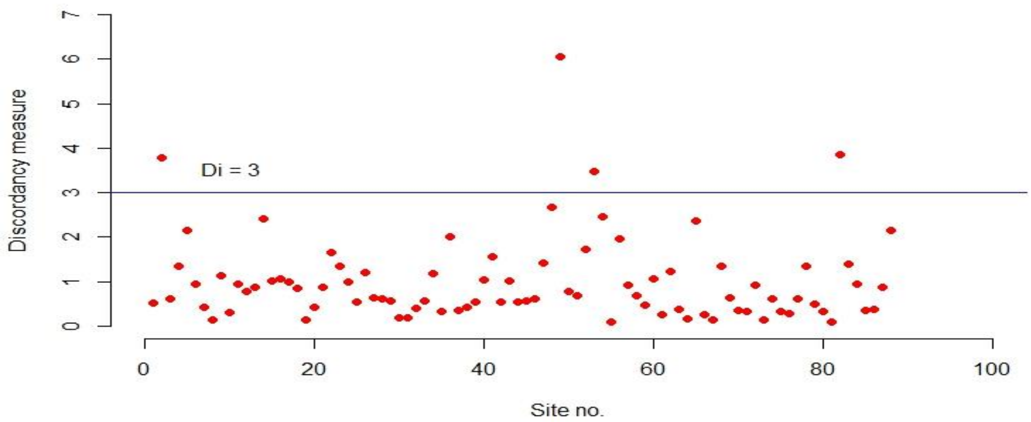

Figure 2 Values of discordancy measure (D_i) for the 88 sites

For further investigation the sites that are found to have $D_i > 3$ are removed, as they can largely affect the H statistic. Next, the H statistics i.e. H_1, H_2 and H_3 are computed for the 88 selected sites and found to be 13.44, 10.06 and 5.96, respectively. Removal of all four sites that have $D_i > 3$, results in H values being equal to 11.74, 8.73 and 4.74, respectively. As there is no significant reduction in the H statistics, these 84 sites (after removal of four discordant sites) are still found to be highly heterogeneous. Furthermore, after removing the four discordant sites, two more sites are found to have $D_i > 3$. These two new discordant sites are 208001 and 222015 having D_i values of 3.37 and 3.69, respectively. Next, these two new discordant sites are removed from the group of 84 sites and D_i and H statistics are computed for the new group of 82 sites. This analysis renders one more discordant site (204026) having D_i of 3.5. H values for the new group of 82 sites are found to be 12.35, 8.91 and 5.01, respectively. It seems the H statistics have increased rather than decreasing after the removal of the discordant sites. Lastly, the new discordant site having D_i of 3.5 is removed from the 82 sites forming a new group of 81 sites and again the D_i and H statistics are computed for the group. This new computation does not result in any more discordant sites. However, H_1, H_2 and H_3 for this latest group of 81 sites are 11.1, 8.31 and 4.83, respectively, which indicates a non-homogeneous group.

Next, each discordant site is removed one at a time from the 88 sites, e.g. site 203002 is first removed and the *H* statistics are computed for the remaining 87 sites; then site 203002 is put back into the group and the next site is removed and the *H* statistics are again computed for the 87 sites as before, and so on. Figure 3 plots the value of discordancy measure for 87 sites. There are four plots (a), (b), (c) and (d), each of them has been derived by removing one discordant site at a time. In each of these plots, it is visible that no matter which discordant site is removed from the group there are still a few discordant sites present in the group.

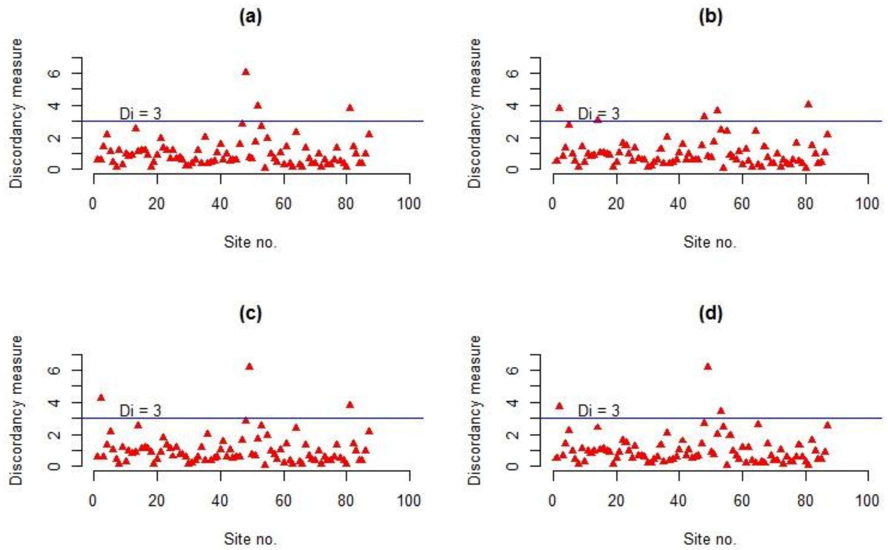

Figure 3 Values of discordancy measure (D_i) for the 87 sites; (a) after removing site 203002, (b) after removing site 222016, (c) after removing site 401015 and (d) after removing site 419029.

Table 1 shows the results from the investigation carried out to determine the effect of individual discordant site on *H* values of the remaining 87 sites. When the first discordant site 203002 is removed from the 88 sites, the other three discordant sites 222016, 401015 and 419029 still remain discordant having D_i values of 6.05, 39.5 and 3.81, respectively and the *H* values have no significant reduction keeping the groups highly heterogeneous. Removal of 222016 (D_i = 6.05) provides four discordant sites and has no significant effect on the *H* statistics. Removal of the other two sites 401015 and 419029 do not generate any better results as shown on the table. Results after removing the two discordant sites one at a time from group of 84 sites (the first four discordant sites are removed from the 88 sites thus rendering a new group with 84 sites) are recorded on table 2. It is visible from the table that, removal of these discordant sites has no significant effect on reduction of *H* values.

Table 1 Sites removed one at a time from first group with 88 sites

Removed site no.	H_1	H_2	H_3
203002	13.43	9.71	5.57
222016	12.16	9.53	5.72
401015	13.32	9.54	5.67
419029	13.31	9.54	5.52

Table 2 Sites removed one at a time from group with 84 sites

Removed site no.	H_1	H_2	H_3
208001	11.22	8.28	4.82
222015	12.22	8.82	4.92

4.2 Heterogeneity analysis for two groups divided from 88 sites based on their drainage division

In the next step of this investigation, all 88 sites are divided into two groups based on their drainage divisions. Group 1 consists of sites that are in drainage division 2 and there are a total of 50 sites in this group; whereas in group 2 there are a total of 38 sites coming from drainage division 4. Both discordancy measure and *H* statistics are computed for these two groups. Application of discordancy measure to both of the groups result in generating two discordant sites from each group i.e. sites 203002 and 222016 from group 1 and sites 401015 and 419029 from group 2. The values of discordancy measure for group 1 and 2 are plotted in figure 4 (a) and (b), respectively.

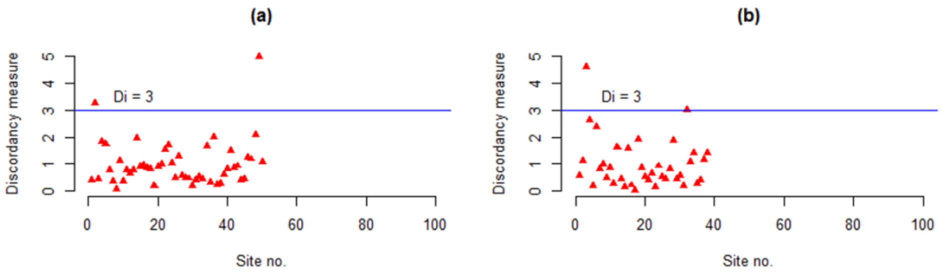

Figure 4 Values of discordancy measure (a) showing 2 discordant sites in group 1 having 50 sites; (b) showing 2 discordant sites in group 2 having 38 sites.

The *H* values for group 1 are found to be 12.81, 9.62 and 6.13, respectively and in case of group 2, these are 5.84, 3.89 and 1.64, respectively. Division of groups according to their drainage divisions has not resulted in any notable reduction in *H* values. Moreover, additional application of the discordancy measure after removing the discordant sites has not developed any remarkable changes. Tables 3 and 4 record the *H* values for group 1 and 2 respectively after removal of the discordant sites.

Table 3 *H* values for group 1 sites obtained from drainage division 2

No. of sites	Sites removed	Removed site names	H_1	H_2	H_3
50	2	203002, 222016	12.81	9.62	6.13
48	1	222015	11.56	9.29	5.83
47	1	208001	11.35	9.09	5.88
46	1	204026	10.66	9.20	6.46
45	1	204056	10.05	8.36	5.85
44	None	N/A	9.96	8.28	5.89

Table 4 *H* values for group 2 sites obtained from drainage division 4

No. of sites	Sites removed	Removed site names	H_1	H_2	H_3
38	2	401015, 419029	5.84	3.89	1.64
36	1	410038	5.55	2.92	0.61
35	1	411001	5.00	2.63	0.59
34	None	N/A	4.64	2.18	0.33

5. CONCLUSION

This study examines identification of homogeneous regions for RFFE in NSW. From the discordancy investigation, it can be stated that although there are few sites as discordant from both the cases (88 sites as a single group and two groups based on drainage division), their removal does not result in any significant reduction in *H* statistics. Also, there is no gross data error discovered for these sites. It can be concluded from these results that perfect homogeneous regions according to the criteria of Hosking and Wallis (1993) cannot be established in NSW. This implies that the index flood method, proposed by Hosking and Wallis (1993) cannot be applied to NSW. Future study will focus on development and testing of an approximate index flood method ignoring the heterogeneity measures.

REFERENCES

Bates BC, Rahman A, Mein RG, Weinmann PE (1998). Climatic and physical factors that influence the homogeneity of regional floods in south-eastern Australia. Water Resources Research, 34, 12, 3369 – 3381.

Castellarin A, Burn, DH, Brath A (2001). Assessing the effectiveness of hydrological similarity measures for flood frequency analysis, Journal of Hydrology, 241, 3–4, 270-285.

Chowdhury JU, Stedinger JR, Lu LH (1991). Goodness-of-fit tests for regional generalized extreme value flood distributions, Water Resources Research, 27, 7, 1765–1776, doi:10.1029/91WR00077.

Dalrymple T (1960). Flood frequency methods, U.S. Geological Survey Water Supply paper, 1543-A.

Fill HD, Stedinger JR (1995). Homogeneity tests based upon Gumbel distribution and a critical appraisal of Dalrymple's test, Journal of Hydrology, 166, 1–2, 81-105.

Greenwood JA, Landwehr JM, Matalas NC, Wallis JR (1979). Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, Water Resources Research, 15, 5, 1049–1054.

Hosking JRM (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics, Journal of the Royal Statistical Society, 52, 1, 105-124.

Hosking JRM, Wallis JR (1993). Some statistics useful in regional frequency analysis, Water Resources Research, 29, 2, 271–281.

Hosking JRM, Wallis JR (1997). Regional frequency analysis: an approach based on L-moments, Cambridge University Press, Cambridge, UK.

Lu LH, Stedinger JR (1992). Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test, Journal of Hydrology, 138, 1–2, 223-245.

Micevski T, Hackelbusch A, Haddad K, Kuczera G, Rahman A (2015). Regionalisation of the parameters of the log-Pearson 3 distribution: a case study for New South Wales, Australia, Hydrological Processes, 29, 2, 250-260.

Noto LV, La Loggia G (2009). Use of L-Moments Approach for Regional Flood Frequency Analysis in Sicily, Italy, Water Resources Management, 23, 11, 2207–2229.

Parida BP, Kachroo RK, Shrestha DB (1998). Regional flood frequency analysis of Mahi-Sabarmati basin (subzone 3-a) using index flood procedure with L-moments, Water Resources Management, 12, 1–12.

Rahman, A., Bates, B.C., Mein, R.G. and Weinmann, P.E. (1999). Regional flood frequency analysis for ungauged basins in south-eastern Australia. Australian Journal of Water Resources. 3(2): 199-207.

Ribeiro-Corréa J, Cavadias GS, Clément B, Rousselle J (1995). Identification of hydrological neighbourhoods using canonical correlation analysis, Journal of Hydrology, 173, 1–4, 71-89.

Saf B (2009). Regional Flood Frequency Analysis Using L-Moments for the West Mediterranean Region of Turkey, Water Resources Management, 23, 3, 531–551.

Wiltshire SE (1986a). Regional flood frequency analysis I: Homogeneity statistics, Hydrological Sciences Journal, 31, 3, 321-333.

Wiltshire SE (1986b). Regional flood frequency analysis II: Multivariate classification of drainage basins in Britain. Hydrological Sciences Journal, 31, 3, 335-346.