Hydrological Sciences–Journal–des Sciences Hydrologiques, **48**(1) February 2003

Regional flood estimation for ungauged basins in Sarawak, Malaysia

YEO HOWE LIM

Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia (currently on study leave at Memorial University of Newfoundland, Canada)

yeohowe@hotmail.com

LEONARD MELVIN LYE

Memorial University of Newfoundland, St. John's, Newfoundland A1B 3X5, Canada

llye@engr.mun.ca

Abstract Design flood estimation is an important task that is required in the planning and design of many civil engineering projects. In this study, the flood records of more than 23 gauged river basins in Sarawak, Malaysia, are examined using an index-flood estimation procedure based on L-moments. Two homogeneous regions were identified and the Generalized Extreme Value and the Generalized Logistic distributions are found to describe the distribution of extreme flood events appropriately within the respective regions. A regional growth curve is subsequently developed for each of the regions. These curves can be used for the estimation of design floods in ungauged basins in Sarawak within the limitations identified for the method. The results presented herein are useful for practicing engineers in Sarawak while the general methodology may be used in any other regions, provided flood records are available.

Key words regional flood frequency; design flood; L-moments; homogeneous region; cluster analysis; index flood; growth curves; Sarawak, Malaysia

Estimation régionale des crues de bassins non jaugés de Sarawak en Malaisie

Résumé L'estimation de crues de projet est une tâche importante, nécessaire dans la planification et le dimensionnement de nombreux ouvrages de génie civil. Dans cette étude, nous analysons les données de crues de plus de 23 bassins versants jaugés à Sarawak en Malaisie grâce à une procédure d'estimation d'un indice de crue basée sur les L-moments. Deux régions homogènes ont été identifiées et il est démontré que les distributions généralisée de valeurs extrêmes et généralisée logistique sont respectivement les plus appropriées pour décrire les distributions des valeurs de crues extrêmes dans ces régions. Une courbe régionale d'accroissement est ensuite élaborée pour chacune de ces régions. Les courbes obtenues peuvent être utilisées, dans les limites identifiées de validité de la méthodologie, pour l'estimation des crues de projet de bassins non jaugés de Sarawak. Les résultats présentés dans cet article sont utiles pour les ingénieurs œuvrant à Sarawak, tandis que la méthodologie générale peut être utilisée dans d'autres régions à condition d'y disposer de données de crues.

Mots clefs fréquence de crue régionale; crue de projet; L-moments; région homogène; analyse de groupe; indice de crue; courbes de croissance; Sarawak, Malaisie

INTRODUCTION

The accurate estimation of design floods remains one of the major challenges for many engineers and planners who are involved in project design where hydrological data and information are limited. This is typical of the case in the eastern State of Sarawak, Malaysia, where many of the rivers remain ungauged. A number of gauged stations in operation also face problems, such as shortness of records, incomplete records, and

Open for discussion until 1 August 2003

inaccuracy of flow rating curves, among others. However, where there is a sufficient number of reliable gauging stations, regionalization can be very helpful in pooling flood data such that design flood estimations can be made at ungauged basins. At present no current regional flood frequency curves are established for the whole of Sarawak. With such circumstances and limitations in mind, regional flood growth curves for the state are developed herein using an approach that is able to minimize the bias due to outliers and shortness of record length. This study strives to provide useful results that can be used by those who need to estimate design floods for non-tidally influenced ungauged basins in Sarawak.

DESIGN FLOOD ESTIMATION PROBLEMS IN SARAWAK

The purpose of design flood estimation is to make predictions on the magnitude of flood discharges at a particular section of a river of interest corresponding to a risk level that is acceptable to the design of structures. The risk is normally taken to be a probability of nonexceedence expressed as a certain return period (*T*-year) or annual recurrence interval (ARI). With sufficient length of flow observations at a particular site of interest, one can make statistical inference on the flood discharges corresponding to various acceptable risk levels.

In the absence of any relevant authoritative guidelines, many practicing engineers in Sarawak are relying on rainfall intensities to generate flood peaks using a simple classical method called the rational method for flood estimations. Despite the fact that the rational method is intended for very small basins, its application in some large basins without any modification is common practice. There are many circumstances where flood estimation procedures applicable to Peninsular Malaysia, published by the Department of Irrigation and Drainage (DID), Sarawak, Malaysia, are adopted for use in Sarawak with the assumption that the basins characteristics are similar. The approach becomes absurd when a particular region in Peninsular Malaysia has to be chosen arbitrarily as similar to Sarawak. The dilemma can be understood in view of the fact that no specific regional flood estimation procedures have been developed for Sarawak. Some rainfall–runoff simulation programs have been used and there are attempts to apply simulation programs using default values of internal parameters without actual comprehensive calibration of regional basins. There are also circumstances where calibrations are done on common flood events and then extrapolated to make estimations for extreme floods.

REGIONAL FLOOD FREQUENCY ANALYSIS

Rationale for regionalization

Regionalization is still a favourite approach in estimating parameters in hydrology compensating for the lack of long hydrological time series and the lack of information. These can be seen in recent papers by Burn & Goel (2000), Cunderlik & Burn (2002), and Pfister *et al*. (2002). It is well accepted that using a regional approach in flood frequency analysis is effective in extending the flood information at a site to sites within a homogeneous region. The extension enables flood quantile estimates for any site in a region to be expressed or inferred in terms of flood data recorded at all gauging sites in that region. Design flood estimations using a regional approach can often be carried out using methods such as the index-flood method and the direct regression on quantiles method. The index-flood method provides an appropriate procedure for statistical flood estimation of extreme events (Dalrymple, 1960), and it has been applied in many countries for flood frequency estimation. *Hydrological Procedure No. 4* published by DID, Malaysia (Heiler & Hong, 1974)**,** and its later versions, also adopted the method. The index-flood method essentially assumes that, within a homogeneous region, the exceedence probability distribution of annual peak discharge is invariant except for a site-specific scaling factor called the index flood. Typically the index flood, Q_m, is taken as the mean of the at-site annual maximum peak discharge series. Robson & Reed (1999) recommended using the median instead of the mean. A relationship can be established between the flood quantile, Q_T, of a site and Q_m with the introduction of a regional growth factor, X_T, that defines the dimensionless frequency distribution common to all sites within a homogeneous region. The relationship is:

$$Q_T = X_T Q_m \qquad (1)$$

A regression of basin characteristic(s) on the index flood can be established based on available information gathered from the gauged sites. Regional growth curves showing the relationship between X_T and the return period, T, can be derived once an appropriate probabilistic distribution has been found within a region with N sites that fits all the gauged flood series, Q_{ij}, where $i = 1, 2, 3, \ldots, N$, $j = 1, 2, 3, \ldots, L_i$, and L_i is the record length at site i. The standardized flood peak:

$$X_{ij} = Q_{ij} / Q_{im} \qquad (2)$$

is used in the estimation of X_T, where Q_{im} is the observed mean or median annual flood at site i. Recent developments in statistical methods (Hosking, 1990; Hosking & Wallis, 1997; Robson & Reed, 1999) has reaffirmed the usage of the index-flood method, and the concept has even been extended to the analysis of site-specific environmental data.

The direct regression-based method is the other significant method still in use (Pandey & Nguyen, 1999). However, the information needed for defining the explanatory variables is often not readily available, especially in developing countries. The index-flood procedure is still the preferred method, as can be seen in many recently published papers, such as those by Fill & Stedinger (1998), Burn & Goel (2000), and Brath *et al.* (2001).

Initial identification of regions by cluster analysis

The first step in regionalization is to delineate regions that are deemed to be homogeneous. The technique of clustering applied to groupings of sites for environmental observation can be found in Fovell & Fovell (1993) and Kalkstein *et al.* (1987). The application of cluster analysis to delineate clusters of basins for regional flood frequency analysis has also been used (Burn & Goel, 2000). Hosking & Wallis (1997) illustrated the use of cluster analysis for regionalization of US annual precipitation totals. Burn & Goel (2000) further improved the technique to include overlapping groups.

By identifying homogeneous regions, one can take advantage of the fact that the more homogeneous a region is, the greater is the gain in using regional instead of at-site estimation. Studies by Dubreuil (1986) on tropical basins show that physical factors such as basin area and slope are the most significant physical explanatory factors for several regression-based flood quantile estimation equations. Various methods of clustering procedures were studied by Kalkstein *et al.* (1987) suggesting that the average linkage method is superior to the centroid and Ward's technique. Hosking & Wallis (1997) also used the average linkage method in the regionalization of annual precipitation totals observed in the United States. The average linkage approach used by Hosking & Wallis (1997) and Burn & Goel (2000) is adopted in this study.

Variables for consideration in clustering are derived by transformation of the site characteristics that are measured at different scales. Appropriate transformation by scaling is necessary to ensure that these factors fall between zero and unity. Initially, all sites are treated as separate clusters. Any two sites that are closest in terms of Euclidean distance are joined. In the next step, either a third site joins the first two, or two other sites join together into a different cluster. This process continues until all clusters are joined into one. A dendrogram can effectively summarize the results of the clustering procedure. The similarity level at any step is the percentage of the minimum distance at that step relative to the maximum inter-observation distance in the data. The decision on how many groups or regions to use, which essentially determines the cut-off level for similarity, is largely heuristic. However, the pattern of how similarity or distance values change from step to step can assist in choosing the final grouping. The step where the values change abruptly may indicate a good point for cutting the dendrogram. Other methods of delineating homogeneous regions are extensively discussed in Robson & Reed (1999), which includes a software package called WINFAP-FEH.

L-moments

Probability weighted moments (PWMs) were introduced by Greenwood *et al.* (1979) as an alternative to conventional moments to minimize the squaring and cubing of observed values. This prevents undue weighting being given to large observed values. However, the method suffers from difficulties in interpretation. Hosking (1990) introduced L-moments, which are linear combinations of PWMs and can be directly interpreted as the measures of scale and shape of probability distributions.

Vogel & Fennessey (1993) highlighted the advantages of L-moments compared with conventional moments. In a wide range of hydrological applications, L-moments provide simple and reasonably efficient estimators of characteristics of hydrological data and of the parameters of a distribution (Stedinger *et al.*, 1992). Ulrych *et al.* (2000) show that probability density functions estimated from L-moments are superior estimates to those obtained from conventional moments and those based on the principle of maximum entropy.

Hosking & Wallis (1993, 1997) refined the index-flood method using the L-moment algorithm and the approach presents an elegant method for flood frequency analysis. It is a general approach that is best suited to environmental data and was developed primarily to analyse regional flood data. Distribution selection has become less subjective using the L-moment approach. The use of the approach in regional flood frequency analysis has gathered momentum over the last decade. An excellent discussion on the use of L-moments in regional flood frequency analysis can be found in Robson & Reed (1999) in addition to Hosking & Wallis (1997).

Discordance

According to Hosking & Wallis (1993, 1997), if a proposed region has N basins, the measure of discordancy, D_i, for basin i is defined as:

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

where $\boldsymbol{u_i}$ is a vector containing the L-moment ratios for basin i, namely the L-CV (t), L-skewness (t_2), and L-kurtosis (t_3), $\bar{\boldsymbol{u}}$ is the unweighted regional average for $\boldsymbol{u_i}$ and $\boldsymbol{A}$ is the matrix of sums of squares and cross products.

Checking on homogeneity

Homogeneity checks using the L-moment approach as proposed by Hosking & Wallis (1993), which were also adopted by Burn & Goel (2000), are based on Monte Carlo simulation. The L-moment ratios t^R, t_3^R and t_4^R for the proposed region are calculated as the sample means weighted proportionally to the record length, l, of i sites. It follows that V, the weighted standard deviation of the at-site sample L-CVs (t_i), is given by:

$$V = \left[\sum_{i=1}^{N} l_i (t_i - t^R)^2 \Big/ \sum_{i=1}^{N} l_i \right]^{1/2} \tag{4}$$

A homogeneity statistic, H, is a measure of the departure of V from similar statistics obtained from the simulation of a large number of realizations for a region with N sites, with μ_v and σ_v as the mean and standard deviation, respectively, of simulated Vs:

$$H = \frac{(V - \mu_v)}{\sigma_v} \tag{5}$$

Hosking & Wallis (1997) suggested that a region is considered to be “acceptably homogeneous” if $H < 1$, “possibly heterogeneous” if $1 \leq H \leq 2$, and “definitely heterogeneous” if $H \geq 2$. Robson & Reed (1999) relaxed the criteria somewhat by suggesting that if $2 < H \leq 4$ a region could be considered as heterogeneous and if $H > 4$ it could be considered as strongly heterogeneous.

L-moment ratios and selection of distribution

The need to fit a suitable distribution cannot be overemphasized, as the flood quantile estimates rely on it alone for extrapolation purposes. Once the appropriate regions have been delineated, the selection of a suitable probability distribution to fit the flood series can be carried out. The L-moment ratios diagram (Vogel & Fennessey, 1993) provides a simple and quick method in selecting the statistical distribution that best fits the data. The point defined by the weighted regional means of τ_4 and τ_3 is plotted on a graph of τ_4 *vs* τ_3 and the theoretical distribution lying closest to the plotted point is selected as an appropriate regional distribution. Peel *et al.* (2001) show by simulations that there are some caveats in using the L-moment ratio diagram, and the use of heterogeneity tests in conjunction with the L-moment ratio diagrams is highlighted.

Goodness-of-fit test

Statistical tests are required to confirm the appropriateness of the chosen distribution and to give a certain degree of confidence in it. A test based on Monte Carlo simulation described by Hosking & Wallis (1993, 1997) is used herein.

For each of the proposed regions, a Kappa distribution with its parameters derived from the fitting of the distribution to the regional average L-moment ratios is used to simulate some large number N_{sim} of realisations for the same region. For each *m*th simulated region, the regional average L-kurtosis t_4^m is calculated. Typical three-parameter distributions are fitted to the sample regional L-moment ratios. For each fitted distribution, the corresponding L-kurtosis, τ_4^{DIST}, is found. The goodness-of-fit measure for each distribution is given by:

$$Z^{DIST} = \left(\tau_4^{DIST} - t_4^R + B_4\right) / \sigma_4 \tag{6}$$

where B_4 is the bias of t_4^R and σ_4 is the standard deviation of t_4^R.

A distribution could be declared as fitting satisfactorily if $|Z^{DIST}| \leq$ 1.64 (Hosking & Wallis, 1993).

REGIONAL FLOOD FREQUENCY ANALYSIS FOR SARAWAK

Study area

The study area, Sarawak, a state of Malaysia on Borneo Island, consists of the western administrative divisions of Kuching, Samarahan, and Sri Aman, less intensely gauged (see Fig. 1) but having sizeable basins to the central and northeastern part of Sarawak, including Batang Rajang, Batang Baram, and Sungai Limbang. Lying almost on the equator, Sarawak has an equatorial-rainforest type of climate and there is no distinct monsoon season. Rainfall varies from about 3000 to over 5500 mm per year, thus endowing the land with vast vegetative growth and many perennial streams and sizeable rivers. One-fifth of Sarawak is composed of alluvial coastal plains covered with low-lying peat swamps.

A large portion of Sarawak is under significant influence of tides. Tidal ranges of up to about 5 or 6 m can be found in the coastal zones of western Sarawak. The estuaries have very mild bed slopes, which enable the tidal influx to reach very far inland. For example, tidal effects can be felt at Kapit on the Rajang River, a distance of over 190 km from the coast. Hence, the tidal interaction zones are considered very

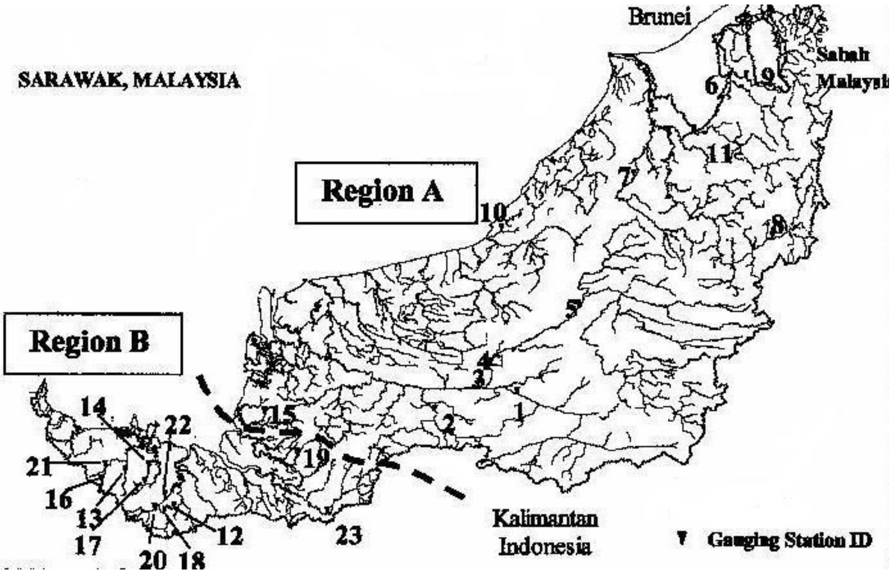


Fig. 1 Gauged river basins in Sarawak and delineated homogeneous regions. (The numbers shown are the site or gauging station ID; the order of numbering follows the listing in Table 1.)

extensive within most of the basins. Flood estimation in tidal interaction zones is discussed in Lim & Lye (2002).

Gauging stations and data screening

All the flowgauging stations within the study area, some having been established in the 1960s, are maintained by the DID. The basin sizes range between 34 and 34 503 km^2. Record lengths are between 9 and 39 years. Figure 1 shows the location of these gauging stations with site details given in Table 1.

Some of the gauged data in the study area were not used because of unfavourable factors, such as a poor discharge rating curve, discontinuity of records, short record length, or influence by tidal backwaters. For gauging stations with disturbed flow regimes, such as Lubok Antu (site 23), the use of records is limited to only those prior to the construction of a dam upstream of the site. The data screening process also involved checking with the discordant statistics D_i (equation (3)). None of the sites that passed the initial screening process were found to be grossly discordant, judging from the computed discordant statistics, D_i, as shown in Table 1. Plotting of L-CV *vs* L-skewness and L-kurtosis *vs* L-skewness was also carried out, as shown in Fig. 2 (for the case "All available sites"). No particular site appeared to be an outlier.

Cluster analysis

Cluster analysis was employed as a first attempt to delineate Sarawak into regions. For this study, the variables considered were: basin area (A), mean annual rainfall (R), return-period storms with duration of 12 h ($T5$, $T10$, $T20$, $T50$), longitude (LO), latitude (LA) and specific discharge (SQ). The variables for consideration in clustering were derived by transformation of the site characteristics shown in Table 1, which

Table 1 Site characteristics and discordant statistics.

Site ID	n	$AREA$ (km^2)	$RAIN$ (mm)	$T5RA$	$T10RA$	$T20RA$	$T50RA$	$LATI$	$LONG$	$SPECQ$	D_i
1	18	9522.0	4343	147	167	186	211	17	36	78.14	0.513
2	17	2273.0	4197	147	167	186	211	16	24	71.71	0.722
3	32	34053.0	3983	133	152	170	193	21	38	73.00	0.757
4	19	21273.0	3883	119	136	153	174	24	39	78.30	0.288
5	29	18261.0	4038	119	136	153	174	25	43	81.45	0.763
6	19	2413.0	2480	150	166	182	202	41	51	66.93	0.102
7	14	2491.0	4512	146	175	202	238	32	43	86.71	1.516
8	34	2690.0	3048	110	129	148	172	34	54	60.81	1.166
9	9	180.0	3346	150	166	182	202	43	51	43.42	2.930
10	16	103.0	3784	187	214	240	274	32	31	47.18	2.237
11	17	3370.0	4194	146	175	202	238	38	50	83.38	2.194
12	34	45.0	3551	148	174	198	230	9	6	55.36	0.353
13	27	440.1	3678	188	225	261	306	12	2	74.60	0.483
14	22	52.5	4107	188	225	261	306	13	4	65.19	1.631
15	15	34.1	3495	138	155	171	192	19	13	58.54	0.978
16	28	124.5	4055	208	255	300	358	13	1	75.98	0.548
17	22	138.3	3903	159	181	202	230	12	4	59.90	0.309
18	17	338.2	3468	141	161	180	205	9	5	65.75	2.147
19	15	320.8	3873	137	156	177	198	15	16	86.38	0.878
20	22	456.1	3192	171	208	243	289	10	3	64.14	0.318
21	28	225.0	4055	208	255	300	358	13	1	70.47	0.742
22	36	950.5	3395	156	185	212	247	10	4	57.97	0.339
23	10	1305.0	3387	132	160	186	221	12	20	88.38	1.086

Note: site characteristics are defined in Table 2. n is the record length and D_i stands for discordant statistics.

Table 2 Transformation of site characteristics.

Site characteristics	Cluster variable
Basin area, $AREA$ (km^2)	$A = AREA/35\,000$
Annual rainfall, $RAIN$ (mm)	$R = RAIN/5000$
T = 5-year 12-h duration storm, $T5RA$ (mm)	$T5 = T5RA/210$
T = 10-year 12-h duration storm, $T10RA$ (mm)	$T10 = T10RA/260$
T = 20-year 12-h duration storm, $T20RA$ (mm)	$T20 = T20RA/310$
T = 50-year 12-h duration storm, $T50RA$ (mm)	$T50 = T50RA/360$
Latitude, $LATI$	$LA = LATI/45$
Longitude, $LONG$	$LO = LONG/55$
Specific discharge, $SPECQ$ (l s^{-1} km^2)	$SQ = SPECQ/100$

Note: Latitude and longitude are indices of the grid squares drawn over the study area, numbered from 1 to 45 and 1 to 55 respectively.

were measured at different scales. Appropriate transformation by scaling, as shown in Table 2, was necessary to ensure that these factors fell between zero and unity. A clustering algorithm using the average linkage method was used.

Clustering amalgamation steps are shown in Table 3. There was a significant change of similarity level from step 20 to 21, dropping 10.66% from 56.96 to 46.3%. The drops in previous steps were within the order of 2–5%. Hence, three clusters could be associated with the cut-off level at 56.96%. Similarly, two main clusters could be

Table 3 Cluster amalgamation steps.

Step	Number of clusters	Similarity level	Drop in Similarity	Distance level	Clusters joined (by ID of sites)	New cluster formed*	Number of sites in new cluster
1	22	96.29	3.71	0.055	16 21	16	2
2	21	92.74	3.55	0.108	12 22	12	2
3	20	91.72	1.02	0.123	4 5	4	2
4	19	91.32	0.40	0.129	12 17	12	3
5	18	90.94	0.38	0.135	13 14	13	2
6	17	89.13	1.81	0.162	19 23	19	2
7	16	87.39	1.74	0.188	12 18	12	4
8	15	86.59	0.80	0.199	7 11	7	2
9	14	85.47	1.12	0.216	13 20	13	3
10	13	82.81	2.66	0.256	2 19	2	3
11	12	79.68	3.13	0.302	6 9	6	2
12	11	78.82	0.86	0.315	12 15	12	5
13	10	78.65	0.17	0.317	13 16	13	5
14	9	75.04	3.61	0.371	6 8	6	3
15	8	72.97	2.07	0.402	2 12	2	8
16	7	71.06	1.91	0.430	1 4	1	3
17	6	65.76	5.30	0.509	6 7	6	5
18	5	63.97	1.79	0.536	1 3	1	4
19	4	59.68	4.29	0.599	2 13	2	13
20	3	56.96	2.72	0.640	6 10	6	6
21	2	46.30	10.66	0.798	1 6	1	10
22	1	33.17	13.13	0.994	1 2	1	23

***** New cluster formed refers to the ID of a site that is first encountered and continued to be used when new members are accepted for entry into the cluster.

identified for the next abrupt change going from step 21 to step 22, with a cut-off level of 46.3%. From the clustering exercise, two major cases (two and three clusters) and two minor cases (one and four cluster(s)) could be identified. Further investigation on the homogeneity of these four cases was carried out using simulations and statistical tests described earlier.

Homogeneity measures

A fitted Kappa distribution for each proposed region was used in generating series of similar record lengths for a similar number of sites in a region. A very large number of realizations, say 10 000, was simulated for each region. The results show that two of the proposed clusters from the four clusters obtained during the initial clustering exercise have very high values of H (see Table 4). By removing sites 8 and 9 from cluster 2, the H value decreased significantly. Sites 8 and 9 are known to have very steep slopes, with basin areas covering the highest peaks in Sarawak. The slope characteristics are unfortunately not available for all the sites, making slope unavailable as a formal factor in the cluster analysis. However, the plots of the moment ratios under the sub-heading “Region A” in Fig. 2 show the outlying nature of these two sites. The next step would be to consider combining clusters to form larger

Table 4 Homogeneity statistics at various steps of delineating regions.

First step		Second step	
Cluster 1 (1, 3, 4, 5)	0.82	Region A (1, 2, 3, 4, 5, 6, 7, 10, 11, 15)	1.56
Cluster 2 (6, 7, 8, 9, 10, 11)	8.59		
Cluster 3 (2, 12, 15, 17, 18, 19, 22, 23)	7.12	Region B (12, 13, 14, 16, 17, 18, 20, 21, 23)	2.02
Cluster 4 (13, 14, 16, 20, 21)	1.31		
		Possible mountainous Region C (8, 9, 19) Site 22 removed	

( ) ID of sites in each group.

Table 5 Final delineation of regions.

ID	Basin	n	l_1	τ	τ_3	τ_4	D_i	V
Region A:								
1	Telok Buing, Rajang	18	1399	0.0418	0.0343	0.0967	0.99	
2	Mukeh, Rajang	17	479	0.1032	0.2101	0.1205	1.42	
3	Kapit, Rajang	32	7188	0.0747	0.1442	0.2476	0.38	
4	Benin, Rajang	19	5842	0.0882	0.0808	0.1795	0.11	
5	Belaga, Rajang	29	5121	0.0780	0.1543	0.2484	0.39	
6	Insungei, Limbang	19	1215	0.1058	0.0723	0.1243	0.78	
7	Long Jegan, Baram	14	452	0.0675	–0.0769	–0.1266	1.74	
10	Sibiu, Kemena	16	30	0.0824	0.0178	0.3612	1.37	
11	Terawan, Baram	17	594	0.0850	–0.1645	0.1327	1.66	
15	Sebatan, Sebatan	15	12	0.0434	0.1229	0.0675	1.17	
Region A weighted average			2872.6	0.0776	0.0737	0.1631		
Total /Value		196						0.01917
Region B:								
12	Bedup, Sadong	34	39.4	0.2233	0.1538	0.0420	0.64	
13	Kpg Git, Sarawak	27	487.1	0.1868	0.0633	–0.0053	0.20	
14	Batu Gong, Samarahan	22	16.6	0.1745	0.2555	–0.0381	1.42	
16	Boring, Sarawak	28	271.8	0.1645	–0.0023	0.0550	0.59	
17	Maang, Samarahan	22	201.8	0.1986	0.1008	0.1586	0.61	
18	Meringgu, Sadong	17	89.5	0.1495	0.3719	0.3324	2.70	
20	Krusen, Sadong	22	423.7	0.1851	0.0569	0.0983	0.34	
21	Buan Bidi, Sarawak	28	348.2	0.2314	0.1332	–0.0226	1.04	
23	Lubok Antu	10	626.8	0.1167	0.1010	–0.0789	1.58	
Region B weighted average			246.8	0.1891	0.1287	0.0565		
Total /Value		210						0.03033

clusters. Clusters 1 and 2 were combined without sites 8 and 9. An acceptable *H* value of 1.56 was still achievable, even including sites 2 and 15 from cluster 3.

Cluster 3 has very high *H* value of 7.12. Site 19, which has a very steep hilly basin, was removed from Region B which is formed by the combination of clusters 3 and 4. Figure 2 shows the plot of L-moment ratios for site 19 in the context of Region B. Site 22 is situated on a wide flood plain, which has many large storage areas reducing the peak discharges. The values in this series are consistently lower than that for Site 20, located upstream. Its removal from the group was desirable as the *H* value was then significantly reduced. Region B was then re-simulated, which gave a favourable

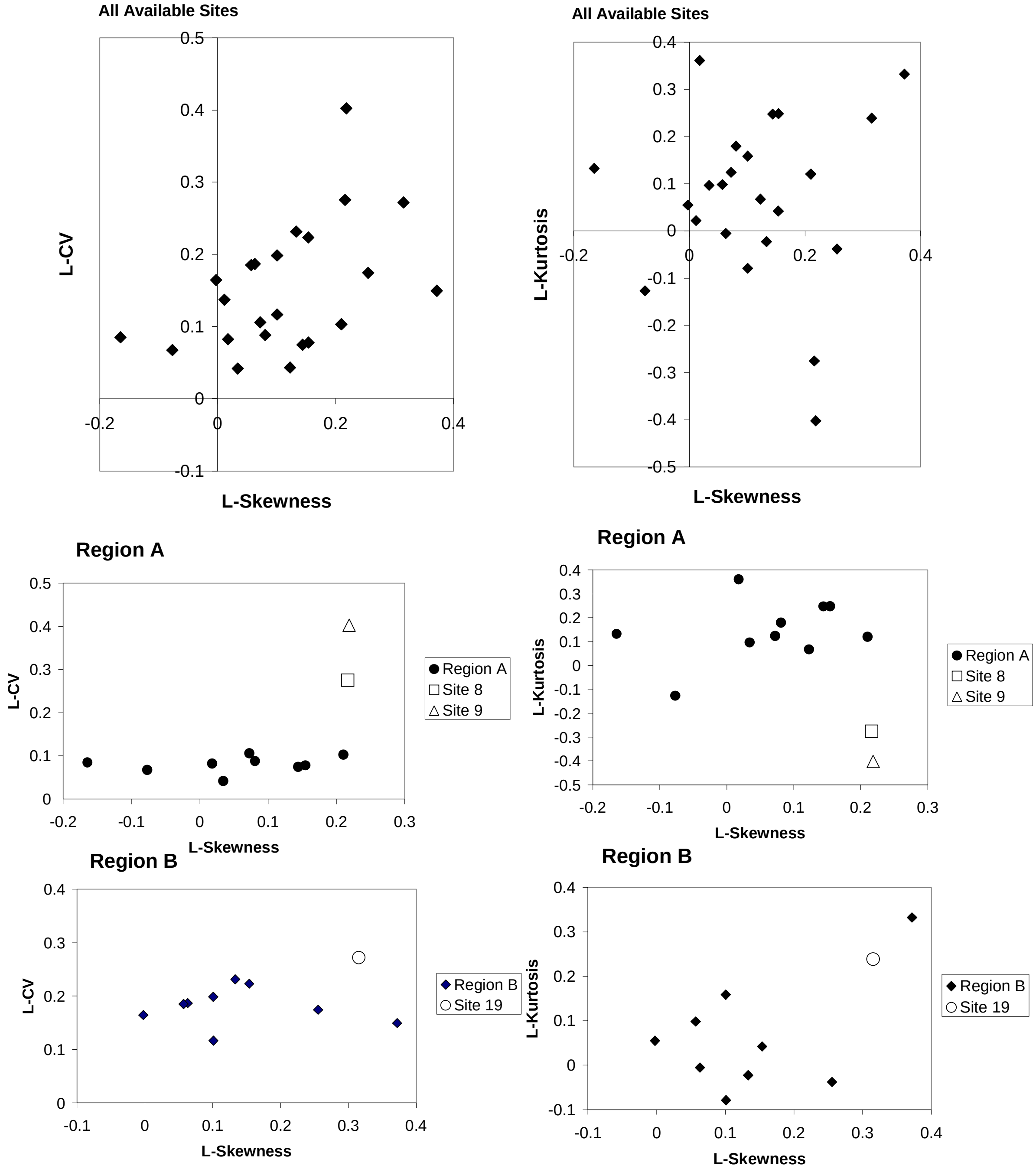


Fig. 2 Plots of L-moment ratios for all the available sites and the two regions.

H value of around 2. The final delineation is shown in Table 5 and Fig. 1. It can be seen from the exercise that, in order to achieve homogeneity, a certain trade-off has to be made on the group size. Detailed discussions on this aspect can be found in Robson & Reed (1999). It should also be noted that, while an attempt has been made to achieve statistical homogeneity, the physical aspects were taken into consideration, at times heuristically, due to the lack of local information such as the slope factor. In fact, it is proposed to group the steep basins under another region. However, the number of sites is limited to only three, well below the accepted minimum of seven sites to form a homogeneous region.

Regional distribution

Using the L-moment diagram as shown in Fig. 3, it was found that Region A, which includes the northeastern part of Sarawak, has a strong preference for the Generalized Logistic distribution, while Region B, covering the western part of Sarawak, has a clear indication of fitting the Generalized Pareto distribution.

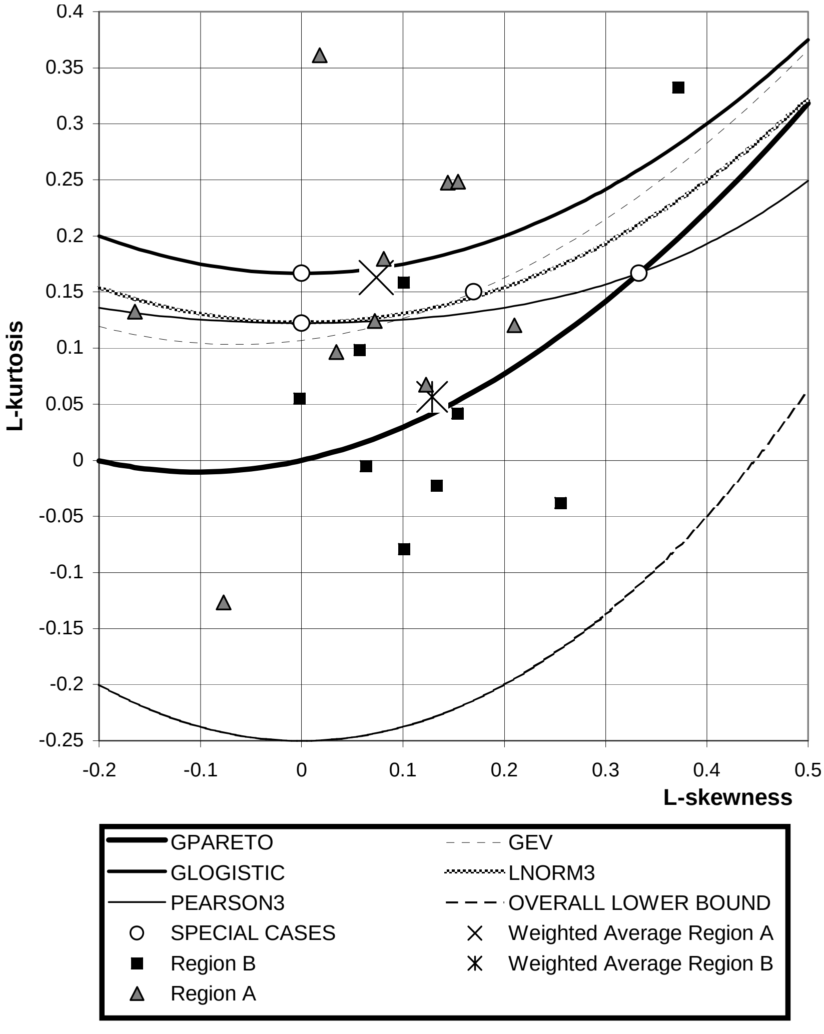


Fig. 3 L-moment ratio diagram for two-region case.

Goodness-of-fit test

The results of the goodness-of-fit test by simulation for the two regions of Sarawak are shown in Table 6. The Generalized Logistic distribution is found to fit well for Region A based on a $|z^{DIST}|$ of 0.259, well below the critical value of 1.64. The lognormal, Pearson Type III, and Generalized Extreme Value (GEV) distributions are also found

Table 6 Results of the goodness-of-fit test for the L-moment approach.

	Generalized Extreme Value	Generalized Pareto	Generalized Logistic	Lognormal	Pearson Type III
Region A:					
τ_4^{DIST}	0.1201	0.0199	0.1712	0.1269	0.1242
$\lvert Z^{DIST}\rvert$	1.52*	5.00	0.259 *	1.281 *	1.375 *
Region B:					
τ_4^{DIST}	0.1355	0.0404	0.1805	0.1356	0.1277
$\lvert Z^{DIST}\rvert$	3.64	0.574 *	5.64	3.67	3.65

* These fits are acceptable as $\lvert Z^{DIST}\rvert \le 1.64$.

to pass the test with $\lvert z^{DIST}\rvert$ of 1.28, 1.38 and 1.52, respectively. For Region B, it can be seen that only the Generalized Pareto distribution fitted to the regional L-moment ratios has $\lvert z^{DIST}\rvert \le 1.64$, at 0.57. The next closest fits are the GEV, Pearson Type III and lognormal distributions, which have $\lvert z^{DIST}\rvert$ of 3.64, 3.65 and 3.67, respectively. It is rare to find the Generalized Pareto distribution fitting well to a regional flood frequency curve. However, the extremely wet equatorial rainforest conditions (e.g. two days in three are rain days), and the flat terrain of Region B may justify the distribution which has an upper bound. The area has very frequent high flows during a year and the specific discharges are high. Another factor that may explain the absence of very extreme floods is that, being on the equator, the region is off the path of all the typhoons which often bring about extreme storm rainfalls in the Philippines and elsewhere in South-east Asia. The findings related to Region B are consistent with the results as reported earlier (Lim & Lye, 2000). Nevertheless, Robson & Reed (1999) gave several reasons not to use the bounded distributions. In this context, the next closest distribution that best fits the regional data is used. This is the GEV distribution, which tends to give more conservative results.

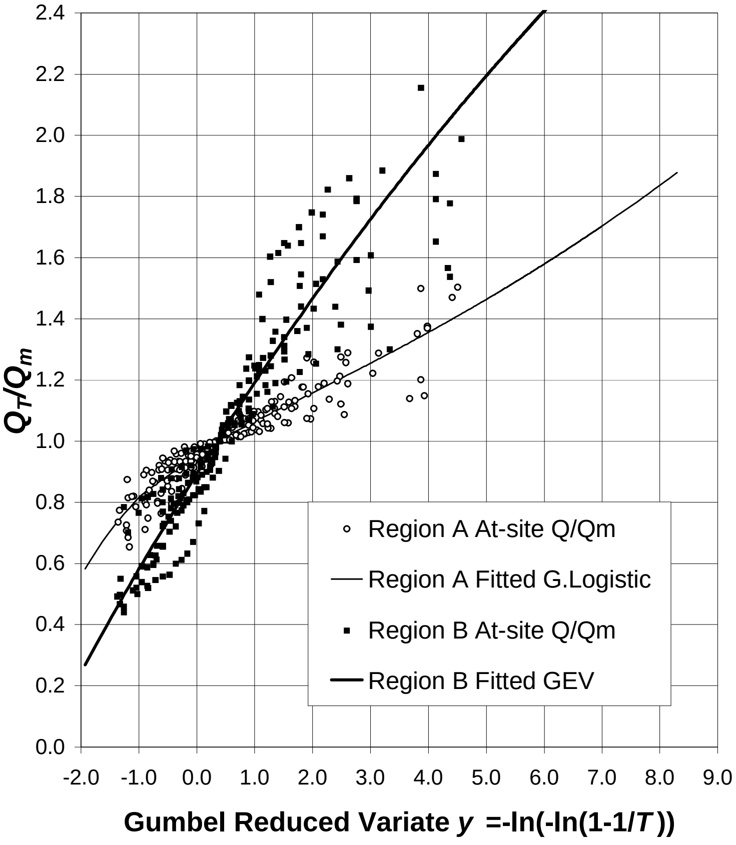
REGIONAL GROWTH CURVES

Once the delineated regions have been shown to be homogeneous and suitable distributions have been identified for the respective regions, regional growth curves can be developed based on the distributions. For this study, region A was associated with the Generalized Logistic distribution while region B was associated with the GEV distribution. Figure 4 shows the proposed regional growth curves fitted for the two regions in Sarawak:

$$\text{Region A: } Q_T/Q_m = 0.9990 - 1.0436\{1 - (T-1)^{0.07367}\} \tag{7}$$

$$\text{Region B: } Q_T/Q_m = 0.8961 + 4.6446\{1 - [-\ln(1 - 1/T)]^{0.06558}\} \tag{8}$$

where Q_T is the flood quantile, Q_m is the at-site median maximum discharge, and T is the return period. For each site, the plotting position formula $(i - 0.35)/n$ was used, where i is the rank of each standardized annual maximum discharge (divided by the median) recorded at the site and n is the sample size of the site. The standardized



Region A: $Q_T/Q_m = 0.9990 - 1.0436\ \{1 - (T-1)^{0.07367}\}$

Region B: $Q_T/Q_m = 0.8961 + 4.6446\ \{1 - [-\ln(1 - 1/T)]^{0.06558}\}$

Fig. 4 Regional growth curves for Sarawak.

annual maximum discharge data for the sites were plotted against the set of plotting positions derived for the site. The plotting positions are related to the Gumbel variate via the return period *T*. The regression of median maximum discharges on basin area can also be developed for each respective region. Ideally, these regression equations should be used when estimating the median annual maximum discharge for any ungauged basin. However, the regression fits are not desirable due to the limited number of basins available. An alternative approach is to use the regression based on all the available basins in the two regions of Sarawak as shown in Fig. 5, with a regression equation of:

$$Q_m = 2.312\ AREA^{0.775} \qquad (R^2 = 0.80) \tag{9}$$

In addition to basin area, basin slope is another important physical variable (Dubreuil, 1986) for tropical regions. The regression is expected to give a better fit to the observations if the data on slopes are available.

AN APPLICATION EXAMPLE

An example is given to illustrate the use of the results presented. Given a bridge site in Sarawak, peak discharges for 25 and 50 years are required for the determination of the

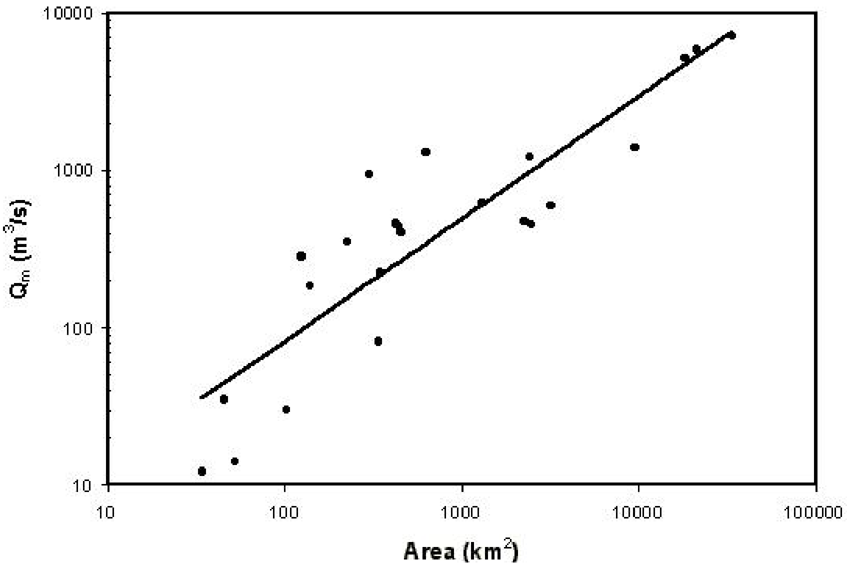


Fig. 5 Regression of median annual peak discharge with basin area for all basins in Sarawak, $Q_m = 2.312AREA^{0.775}$ ($R^2 = 0.80$).

bridge deck height and for bank protection design. As the site is remote, no gauging record is found for this river. The basin area upstream of the bridge site is obtained from a topographic map as 265 km^2. With reference to the regional map of Sarawak (Fig. 1), the basin is initially identified to be in Region A. Using the full factors as listed in Table 1, the basin is confirmed to be within Region A. Hence, from Fig. 4 (or by equation (7)) the growth factors of 1.217 and 1.259 are obtained for $T = 25$ and $T = 50$ years, respectively. From Fig. 5 (or the regression equation (9)), Q_m corresponding to a basin area of 265 km^2 is 174.58 m^3 s^{-1}. Hence, the flood quantiles Q_{25} and Q_{50} are estimated at 212.5 and 219.8 m^3 s^{-1}, respectively.

LIMITATIONS

In some major basins of Sarawak, representative gauging stations are absent. The application of the regional growth curves to those areas is not recommended. For a state occupying almost half of Malaysia's land mass, the number of gauging stations in Sarawak is still below the standard density in a streamgauging network as recommended by WMO (1981). Hence certain large areas, such as the Kemena River Basin area (around Site 10), do not have any credible representative gauged station. The uncertainty in estimation of flood quantiles would be high in those outlying areas.

CONCLUSION

A practical regional flood estimation procedure for Sarawak, Malaysia, has been presented, employing the techniques in regional flood frequency analysis. This is a big step forward in the local context. Although it is limited, in terms of both the spatial coverage of the gauged basins and the record length, the identification of the two homogeneous regions is viable after testing rigorously for homogeneity in a statistical sense. The selection of the appropriate frequency distribution was also conducted based on an appropriate statistical test. With due respect to the limitations as discussed, the regional flood frequency results can be applied to non-tidally influenced ungauged basins in Sarawak lying within or in proximity of an identified homogeneous region. The methodology used in this study can be adopted for other regions provided that sufficient flood records are available.

Acknowledgements Most of the flow data used in this study were supplied by the Department of Irrigation and Drainage, Sarawak, Malaysia. The useful comments of two anonymous reviewers are gratefully acknowledged.

REFERENCES

Brath, A., Castellarin, A., Franchini, M. & Galeati, G. (2001) Estimating the index flood using indirect methods. *Hydrol. Sci. J.* **46**(3), 399–418.

Burn, D. H. & Goel, N. K. (2000) The formation of groups for regional flood frequency analysis. *Hydrol. Sci. J.* **45**(1), 97–112.

Cunderlik, J. M. & Burn, D. H. (2002) The use of flood regime information in regional flood frequency analysis. *Hydrol. Sci. J.* **47**(1), 77–92.

Dalrymple, T. (1960) Flood frequency analysis. *US Geol. Survey Water-Supply Paper no. 1543–A.*

Dubreuil, P. L. (1986) Review of relationships between geophysical factors and hydrological characteristics in the tropics. *J. Hydrol.* **87**, 201–222.

Fill, H. D. & Stedinger, J. R. (1998) Using regional regression within index flood procedure and an empirical Bayesian estimator. *J. Hydrol.* **210**, 128–145.

Fovell, R. G. & Fovell, M. Y. C. (1993) Climatic zones of the conterminous United States defined using cluster analysis. *J. Climate* **6**, 2103–2135.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. & Wallis, J. R. (1979) Probability weighted moments: distribution and relation to parameters of several distributions expressable in inverse form. *Water Resour. Res.* **15**(5), 1049–1054.

Heiler, T. D. & Hong, C. H. (1974) Magnitude and frequency of floods in Peninsula Malaysia. Hydrological Procedure no. 4, Ministry of Agriculture and Fisheries, Malaysia.

Hosking, J. R. M. (1990) L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Roy. Statist. Soc. Sec. B* **52**(1), 105–124.

Hosking, J. R. M. & Wallis, J. R.(1993) Some statistics useful in regional frequency analysis. *Water Resour. Res.* **29**(2), 271–281.

Hosking, J. R. M. Wallis, J. R. (1997) *Regional Frequency Analysis—An Approach Based on L-moments*. Cambridge University Press, Cambridge, UK.

Kalkstein, L. S., Tan, G. & Skindlov, J. A. (1987) An evaluation of three clustering procedures for use in synoptic climatological classification. *J. Am. Met. Soc.* **26**, 717–730.

Lim, Y. H. & Lye, L. M. (2000) Selection of a regional flood frequency distribution for Western Sarawak, East Malaysia. In: *Proc. CSCE 28th Annual Conf.* (London, Ontario, 7–10 June 2000) CD-ROM, Canadian Society of Civil Engineering, Montreal, Quebec.

Lim, Y. H. & Lye, L. M. (2002) Design flood estimation methods in tidal interaction zones. In: *Water Engineering and Geomatics* (Proc. Second World Engineering Congress - Engineering Innovation and Substainanility: Global Challenges and Issues, Kuching, Sarawak, Malaysia, 22–25 July 2002) (ed. by M.J.M.M. Noor *et al.*), 1–6. Universiti Putra Malaysia Press, Serdang, Selangor Darul Ehsan, Malaysia.

Pandey, G. R. & Nguyen, V.-T.-V (1999) A comparative study of regression based methods in regional flood frequency analysis. *J. Hydrol.* **225**, 92–101.

Pearson, C. P. (1995) Regional frequency analysis of low flows in New Zealand rivers. *J. Hydrol. (NZ)* **33**(2), 94–122.

Peel, M. C., Wang, Q. J., Vogel, R. M. & McMahon, T. A. (2001) The utility of L-moment ratio diagrams for selecting a regional probability distribution. *Hydrol. Sci. J.* **46**(1), 147–155.

Pfister L, Iffly, J.-F. & Hoffmann, L. (2002) Use of regionalized stormflow coefficients with a view to hydroclimatological hazard mapping. *Hydrol. Sci. J.* **47**(3), 479-491.

Robson, A. J. & Reed, D. W. (1999) *Flood Estimation Handbook*, vol. 3: *Statistical Procedures for Flood Frequency Estimation*. Institute of Hydrology, Wallingford, UK.

Stedinger, J. R., Vogel, R. M. & Foufoula-Georgiou, E. (1992) Frequency analysis of extreme events. In: *Handbook of Hydrology* (ed. by D. R. Maidment), Chapter 18. McGraw-Hill, New York, USA.

Ulrych, T. J., Velis, D. R., Woodbury, A. D. & Sacchi, M. D. (2000) L-moments and C-moments. *Stochast. Environ. Res. and Risk Assessment* **14**, 50–68.

Vogel, R. M. & Fennessey, N. M. (1993) L Moment diagrams should replace product moment diagrams. *Water Resour. Res.* **29**(6), 1745–1752.

WMO (1981) *Guide To Hydrological Practices*, vol. I (fourth edn), WMO Report no. 168. World Meteorological Organization, Geneva, Switzerland.

Received 6 February 2002; accepted 22 October 2002