Stochastic Hydrology and Hydraulics 12 (1998) 65–82

Regionalization of floods in New Brunswick (Canada)

N. El-Jabi, F. Ashkar, S. Hebabi

Abstract This study uses the method of peaks over threshold (P.O.T.) to estimate the flood flow quantiles for a number of hydrometric stations in the province of New Brunswick, Canada. The peak values exceeding the base level (threshold), or 'exceedances', are fitted by a generalized Pareto distribution. It is known that under the assumption of Poisson process arrival for flood exceedances, the P.O.T. model leads to a generalized extreme value distribution (GEV) for yearly maximum discharge values. The P.O.T. model can then be applied to calculate the quantiles X_T corresponding to different return periods T, in years. A regionalization of floods in New Brunswick, which consists of dividing the province into 'homogeneous regions', is performed using the method of the 'region of influence'. The 100-year flood is subsequently estimated using a regionally estimated value of the shape parameter of the generalized Pareto distribution and a regression of the 100-year flood on the drainage area. The jackknife sampling method is then used to contrast the regional results with the values estimated at site. The variability of these results is presented in box-plot form.

Key words: Flood flow, threshold, generalized Pareto, Poisson

1 Introduction

The many uses of water resources include irrigation, recreational activities, sport and commercial fishing. Among the most important uses is the production of energy. Typically, this use requires substantial economic and social investments

Received: June 1, 1997

N. El-Jabi, S. Hebabi
School of Engineering, Université de Moncton,
NB, E1A 3E9, Canada

F. Ashkar
Department of Mathematics and Statistics,
Université de Moncton, NB, E1A 3E9, Canada

Correspondence to: F. Ashkar, Fax: (1) 506-858 4396,
e-mail: ashkarf@umoncton.ca

The financial support for this study, provided by the Natural Sciences and Engineering Research Council of Canada, is gratefully acknowledged. The authors would like to thank the referees for their useful suggestions, and Mr. Daniel Bastarache for his kind help in the preparation of this manuscript.

associated with the construction of large hydro-technical structures. These structures and other smaller works (e.g. bridges and culverts), if inadequately designed, may be subjected to natural destructive phenomena, such as flooding.

In practice, absolute control of the destruction due to flooding may not be attainable. Damages and costs can be minimized by estimating the flood flow conditions that may affect a structure during its useful life. A comprehensive knowledge of streamflow is important in the analysis of the frequency of flooding. Flood frequency analysis uses measurements taken during long recording periods at hydrometric station sites, and a regional analysis can be performed to classify sites with similar behaviour characteristics to aid in the extrapolation of information to ungauged sites. Classification of 'homogeneous regions' aids in the prediction of extreme streamflow conditions used in engineering design.

Regionalization has long been used as instrument to facilitate the extrapolation of data from gauged to ungauged sites. One of the pioneers of regionalization, Dalrymple (1960), developed the method of 'flood indexing', still in wide use today. The procedure of regionalization is becoming a standard pre-requisite for adequate solutions to innumerable problems related to the management of hydric networks (Simmers 1984).

The majority of regional flood frequency estimation studies reported in the literature have been done using the annual flood series approach, and very few have used the P.O.T. method. The objective of this study is to present a regionalization approach of quantiles with the aim to estimate the extreme flow evaluated by the method of P.O.T. A numeric application of the regionalization of flood parameters is performed on 53 hydrometric stations in New Brunswick (Canada). In this application, the generalized Pareto distribution is used to fit flood peak values exceeding the threshold, but other distributions can also be used within the P.O.T. framework. The regionalization is done first on the coefficient of skewness of the generalized Pareto distribution, and then a direct regression of the 100-year flood on the drainage area is performed. The jackknife sampling procedure is used both to contrast these two regionalization methods, and to compare regional values with those estimated at-site. The variability and bias of the results is facilitated by the use of box plots.

2
Background

Before analyzing the background of hydrologic data, it is opportune to present some definitions of a homogeneous region. Tasker (1982) used cluster analysis to define homogeneous regions. He describes this analysis as being the exploration and organisation of data to group the recording stations that are similar. Simmers (1984), among others, studied the problem of defining criteria for homogeneous regions. He made an attempt to rationalise the steps involved in the practical application of regionalization, using a multidisciplinary systematic approach.

To define homogeneous regions, Acreman and Sinclair (1986) base their work on multivariate analysis of the characteristics of the basin. Like Tasker (1982), they use cluster analysis to form basin groups physically comparable. Acreman and Sinclair assign each basin a probability of association to each group. Basins are transferred between groups if greater similarities to other groups are perceived. The process continues until the maximum resemblance has been achieved.

Burn (1989) also applied a cluster analysis to form homogeneous regions based on different characteristics of the basin. He applied the method to river data in south Manitoba, Canada.

To define homogeneous regions, Burn (1990 a,b), subsequently, used the 'region of influence' approach which allocates to each site a unique set of similar stations. In this method, as with previous methods proposed by Wiltshire (1986 a,b,c), the definition of geographic boundaries between regions is unnecessary. Each region is composed of several sites measured by proximity to the control or target station, this proximity being given by Euclidean distance. All stations with a distance from the control station greater than that of a predetermined threshold value are excluded from the region of influence. The selection of the threshold value requires some prudence.

3
Flood flow evaluation by the P.O.T. method

Among the basic studies presenting the P.O.T. approach to flood frequency analysis are those by Zelenhasic (1970), and Rousselle (1972). The P.O.T. approach is based on the analysis of flood 'exceedances', which are peak discharges above a base level (threshold), Q_B. Both studies by Zelenhasic and by Rousselle are based on extreme value theory as presented by Todorovic (1970). The general P.O.T. model divides the time axis into successive intervals $(T_{k-l}, T_k]$ ('seasons') and assumes that flood exceedances are independent and identically distributed (iid) within each season.

The foregoing and some subsequent references (e.g., Ashkar et al. 1991, Rosbjerg et al. 1992, Rasmussen et al. 1994) have also found it reasonable to assume a Poisson process arrival for flood exceedances. The cumulative distribution function (cdf) of the maximal exceedance $\chi(t)$ in an arbitrary time period $(0, t]$ under this assumption is (Rousselle 1972):

$$F_t(x) = \exp\left\{-\sum_{i=1}^{k-1}[\Lambda(T_i) - \Lambda(T_{i-1})][1 - H_i(x)]\right\} \times \exp\{-[\Lambda(t) - \Lambda(T_{k-1})][1 - H_k(x)]\} \quad (1)$$

where t is k 'seasons' away from the origin (i.e., $t \in (T_{k-l}, T_k]), \Lambda(t)$ is the average number of exceedances $\xi \in (O, t]$, and $H(x)$ is the cdf of the exceedance values, i.e.

$$H(x) = P\{\xi \leq x\} \quad (2)$$

This cdf $H(x)$ is represented by one of the common statistical distributions (exponential, lognormal, gamma, ...) dependent on the hydro-meteorologic characteristics of the region. For example, in many flood cases the exponential cdf adequately describes the exceedances; its cdf is

$$H(x) = 1 - e^{-\beta x} \quad \beta > 0 \quad (3)$$

with $\beta = \{E[\xi]\}^{-l}$, where E is the expression for mathematical expectation. Assuming this exponential form for $H(x)$, two special cases of eq. (1) have been considered in practice:

(1) by taking the time interval (0,t] to be equal to 1 year, and if the exceedances are iid throughout the year, then (Zelenhasic, 1970):

$$F_t(x) = \exp\{-\Lambda(t)e^{-\beta x}\} \quad (4)$$

where $\Lambda(t)$ is the average number of exceedances per year, and β is the parameter of the exponential distribution;

(2) if the exceedances are only iid within seasons (and assuming the year interval (0,t] to be divided into the four usual seasons), then (Rousselle 1972):

$$F_t(x) = \exp\{-\Lambda(T_1)e^{-\beta_1 x} - [\Lambda(T_2) - \Lambda(T_1)]e^{-\beta_2 x} - [\Lambda(T_3) - \Lambda(T_2)]e^{-\beta_3 x} - [\Lambda(T_4) - \Lambda(T_3)]e^{-\beta_4 x}\} \quad (5)$$

where $\Lambda(T_1), [\Lambda(T_2) - \Lambda(T_1)], \ldots, [\Lambda(T_4) - \Lambda(T_3)]$ represent the average number of exceedances, and $\beta_1, \ldots, \beta_4$ are the parameters of the exponential distribution, for each season.

In this study, the exceedances are considered iid throughout the 'year', where the term 'year' is defined later. We also take the cdf $H(x)$ to be that of the generalized Pareto distribution:

$$\begin{aligned} H(x) &= 1 - \left(1 - k\frac{x}{\alpha}\right)^{1/k} \quad k \neq 0 \\ &= 1 - \exp\left(\frac{-x}{\alpha}\right) \quad k = 0 \end{aligned} \quad (6)$$

It is known that while conserving the independence hypothesis among floods, and under the assumption of a Poisson process arrival for flood exceedances, $\chi(t)$ follows a generalized extreme value distribution (GEV) with the cdf (Rosbjerg et al. 1992):

$$\begin{aligned} F_t(x) &= \exp\left\{-\lambda\left[1 - \frac{kx}{\alpha}\right]^{1/k}\right\} \quad k \neq 0 \\ &= \exp\left[-\lambda e^{-x/\alpha}\right] \quad k = 0 \end{aligned} \quad (7)$$

where $\lambda = \Lambda(t), t =$ one 'year', is the average number of exceedances per 'year'. Note that the shape parameter k of the generalized Pareto distribution is the same as that of the generalized extreme value distribution. In hydrologic applications, the values of interest of the parameter k range between -0.5 and 0.5 (Hosking and Wallis 1987).

An estimator of the quantile X_T of the generalized Pareto distribution (flow corresponding to a return period T) is given by:

$$\begin{aligned} X_T &= Q_B + \frac{\hat{\alpha}}{\hat{k}}\left[1 - \left(\frac{1}{\hat{\lambda}}\ln\frac{T}{T-1}\right)^{\hat{k}}\right] \quad k \neq 0 \\ &= Q_B + \hat{\alpha}(\ln\hat{\lambda} + C) \quad k = 0 \end{aligned} \quad (8)$$

where $\hat{\alpha}, \hat{k}$, and $\hat{\lambda}$ are estimates of α, k and λ, respectively. Q_B is the base level (threshold) and C = –ln $[\ln(T/(T-1))]$.

4
Regional analysis

The region of influence approach permits the use of the data from several recording stations to estimate the at-site extreme values for each station in the group. An important property of this approach is the elimination of distinct boundaries between the regions. In effect, extreme values of a site are estimated using data from 'similar' stations (i.e. the region of 'influence'). The region of influence approach consists of determining the proximity of each station to the others (measured using the weighted Euclidean distance in P-dimensional space, where P represents the number of attributes used in the identification of the similarity of the stations). This distance is given by

$$D_{ij} = \left[\sum_{m=1}^{P} W_m (x_m^i - x_m^j)^2 \right]^{1/2} \qquad (9)$$

where D_{ij}: weighted distance between site i and station j
W_m: weight applied to the attribute m to reflect its relative importance
P: number of attributes
x_m^i: standardized value of the measure of attribute m for site i

This standardization is applied to eliminate the measurement units and to avoid the introduction of bias due to scale differences of the attributes.

The success of the method is entirely dependent on the selection of attributes. The choice is based on the availability of site data and scientific judgment. For gauged sites a correlation between extreme flows and potential attributes facilitates the selection. In the case of several attributes a multi-dimensional analysis is performed (Burn 1990 a,b). Two categories of attributes may be used in the measure of the distance:

(a)– attributes based on the physical properties of the drainage basin (e.g. latitude, longitude, soil type, lakes and swamps);
(b)– attributes based on the statistical parameters determined for each station (e.g. the coefficient of variation, the coefficient of skewness).

In both cases, the attributes are related to the extreme flows. The stations to include in the region of influence of site i are those belonging to the set

$$I_i = \{j : D_{ij} \leq S_i\} \qquad (10)$$

where I_i: is the group of stations for site i
S_i: is the threshold value for site i

A careful choice of the threshold value S_i is important because an overly large value can lead to an unnecessary increase in the number of stations in the region, which is translated into a low degree of homogeneity among the stations (high 'quantity' but low 'quantity' of information). Inversely, a threshold value that is not sufficiently high, increases homogeneity among the sites, but information is lost due to a decrease in the number of stations in the region.

Upper and lower limits on the threshold value S_i can be imposed to obtain a better control over the number of stations in the region of influence for station $i (j \in I_i)$. This can be done as follows:

(1) *Take* $S_i = S_L$ *if* $NS_i \geq NCS$ (11)

or

(2) $Take\ S_i = S_L + (S_U - S_L)\left(\frac{NCS - NS_t}{NCS}\right)\ if\ NS_i < NCS$ (12)

where S_i: threshold value for site i
S_L: lower threshold limit
S_U: upper threshold limit
NS_i: number of stations in the region of influence of site i
NCS: target number of stations

Equation (12) is used to increase the number of stations $j \in I_i$ in the case where this number is insufficient. To reflect the relative proximity of the site in relation to each station in the region of influence, the following general function can be used:

$$f_{ij} = g(D_{ij}, \theta) \quad \forall j \in I_i \tag{13}$$

$$f_{ij} = 0 \quad \forall j \notin I_i \tag{14}$$

where f_{ij}: the weight for station $j \in I_i$
$g()$: the specific relationship that defines this weight as a function of D_{ij}, with θ a parameter.

A specific weighting function f_{ij} used by Brun (1990 a) is:

$$f_{ij} = 1 - \left(\frac{D_{ij}}{H}\right)^n \tag{15}$$

where H: a parameter
n: a positive constant

The parameter H assigns a weight to the stations, thus if $H = D_{ij}$, a station j situated at the threshold value does not contribute to the flood estimation at site i. The other parameter n determines the rate of decrease of the weighted values.

Finally, the regional estimation of flow corresponding to recurrence period T, for station i, is done by the following procedure:

1. A statistical method is used to estimate the shape parameter (k) of the generalized Pareto distribution at each station (j) in the region of influence of station $i (j \in I_i)$. In this study, the estimation is done by the method of moments (Hosking and Wallis 1987), as follows:

   $$\hat{k} - \frac{1}{2}\left(\frac{\bar{x}^2}{s^2} - 1\right) \tag{16}$$

   where $\bar{x}$ and s^2 are the mean and variance of the exceedances at the station (j) under consideration.
2. The coefficient of skewness C_s of the generalized Pareto distribution, which is a dimensionless function of the shape parameter k, is calculated for each station $j \in I_i$ by

   $$C_{s_j} = 2(1 - k_j)(1 + 2k_j))^{1/2}/(1 + 3k_j) \tag{17}$$

   (Note that by using eqs. 16 and 17, C_s is estimated without having to use the moment of order 3).

3. The regional C_s value for stations $j \in I_i$ (denoted by C_s*_i) is calculated by

$$C_s*_i = \sum_{j \in I_i} N_j C_{sj} f_{ij} \Big/ \sum_{j \in I_i} N_j f_{ij} \quad (18)$$

where N_j: the number of recording years for station j
f_{ij}: weight function defined by Equation (15)
C_{sj}: coefficient of skewness for station j defined by Equation (17).

4. The regional estimate of the flow corresponding to recurrence period T, for station i (denoted by $X*_{Ti}$) is calculated by

$$X*_{Ti} = Q_{Bi} + \frac{\hat{\alpha}_i}{k*_i}\left[1 - \left(\frac{1}{\hat{\lambda}_i}\ln\frac{T}{T-1}\right)^{k*_i}\right] \quad (19)$$

where Q_{Bi}: the base level (threshold) at station i
k_i^*: calculated from C_s^{*i} by eq. (17), which is the basic equation relating k to C_s.

5 Case study

The method of peaks over threshold was used to perform flood frequency analysis in New Brunswick using data from 53 hydrometric stations (fig. 1). Each station was subjected to a detailed analysis in order to estimate the recurring flows for different return periods. This analysis consisted of verifying if the exceedance arrivals could be represented by a Poisson process, and if the exceedance magnitudes could be represented by a generalized Pareto distribution. In order to avoid dealing the mixtures of flood exceedance populations, only floods within the 'humid season' were considered. For the region under study, this season was taken to cover the six months of January to June. The 'year' was considered to be composed of these six months for the purpose of applying the P.O.T. model. Flood exceedances were considered iid throughout this 'year' interval. The choice of threshold Q_B should not be done independently of the choice of distribution type to represent the exceedance magnitudes. Careful examination of the data at each of the 53 hydrometric stations (fig. 1) was therefore needed to allow the choice of threshold level that was appropriate for a generalized Pareto fit. Statistical tests were executed to assess the goodness of fit of this distribution to the data. Finally, quantiles are estimated for different periods of return from this *local* (at-site) analysis of each flood record (i.e. no transformation of information between sites is carried out up to this point). Table 1 gives the results.

To improve the estimation of the quantiles, a regional study can be performed to form potential homogeneous regions taking into consideration regionally recorded flood flows and physiographic characteristics of basins. A regional study of the 100-year flow by the P.O.T. method was performed and tested at eight stations considered to be representative of the sites in New Brunswick. This flow was calculated from the regionalization of a shape descriptor of the generalized Pareto density. As mentioned earlier, the shape descriptor used was the coefficient of skewness, C_s, which is a dimensionless function of the shape parameter k (eq. 17). For the purpose of comparison, the same calculations were also executed using 'direct regression' based on the relation between characteristics of basins and quantiles (100-year flow). An advantage of this method, which needs to be investigated more fully in future studies, is the ability to apply different distributors to stations.

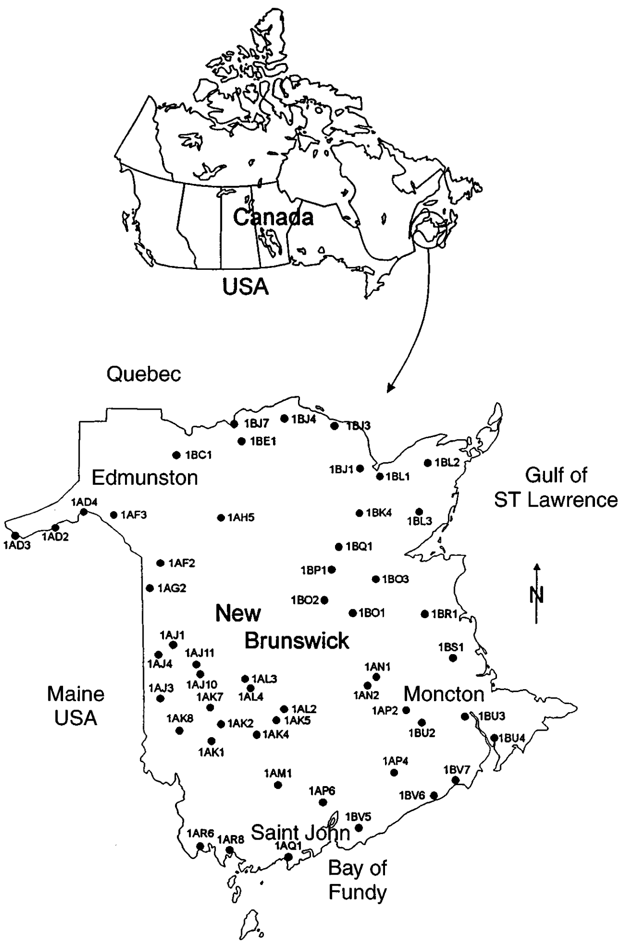


Fig. 1. Hydrometric stations in New Brunswick

The attributes introduced for the calculation of proximity of each station to others were chosen as a function of their relation to river flows and their data availability. In this study, the longitude and latitude were used as physical attributes reflecting the position of the stations. This choice may be justified by the fact that stations that are physically close are subject to the same meteorologic effects and receiving for example similar quantities of precipitation. Two other attributes chosen are directly related to extreme flows. These include the coefficient of variation (C_v) of exceedances (or exceedances + threshold Q_B) and a variable equal to average flow $\bar{Q}$ divided by basin area (A). The chosen attributes are presented in table 2. The proximity of each station in relation to other stations is calculated using eq. (9).

For each station the elements of the region of influence, the minimum number of stations desirable in each region and the threshold values are calculated. The threshold value is determined from the distances D_{ij} (eq. 9), which controls this number and sets the membership of a station to a region. The number of stations was fixed at ten, representing approximately a fifth of the total number of stations. The threshold values in eqs. (11) and (12) are determined by a compromise

Table 1. T-year flow (m^3/s) for different return periods

Station	Return period (years)				
	2	10	25	80	100
AD002	1804.7	3226.4	3646.6	3876.6	4083.3
AD003	180	326.9	383.3	415.5	441
AD004	1946.8	4181.2	4674	5283.5	5500.6
AF002	2582.3	4912	5677.8	6122.8	6462.2
AF003	196.9	358.3	414.9	449.3	478.1
AG002	26.2	45.9	56.3	64.2	72.3
AH005	34	63.5	78.8	90.4	102.1
AJ001	3842	7405	8614.8	9330.8	9918
AJ003	205.9	378.3	472.7	549.3	630
AJ004	88	165.7	196.6	220.7	240.8
AJ010	68.6	109.5	124.6	133.9	141.8
AJ011	27.7	52.9	69	82.6	97.9
AK001	29.8	58.9	71.1	79.3	86.7
AK002	4418.7	7104.3	7767.9	8096.6	8328.9
AK004	5392.6	9072.9	10264.3	10950.6	11501
AK005	5.5	9.3	12	13.7	15.4
AK007	46.9	79.7	95.4	106.6	117.5
AK008	53.5	91.1	106.9	117.6	127.3
AL002	272.7	496.3	606.9	688.2	766.1
AL003	1.7	3.9	5.5	6.9	8.5
AL004	1.1	2.1	2.8	3.4	4.1
AM001	96.8	192.2	253.6	306.3	365.5
AN001	7.7	11	11.7	12	12.3
AN002	189.1	244.6	264.9	277.4	263
AP002	112.8	189.7	218.9	237.4	253.4
AP004	169.1	313.4	396	462.1	532
AP006	74.4	126.9	151.9	169.9	187.2
AQ001	45.3	107.5	154.4	196.7	252.6
AR006	21.6	35.8	40.9	44	46.6
AR008	8.5	18.2	25.3	32	40
BC001	521	836.7	1091.6	1188.2	1271.1
BE001	256.8	529.2	634.7	702.1	760.9
BJ001	86.9	108.7	128.8	143.3	157.4
BJ003	106.1	152.9	169.6	179.6	188.1
BJ004	24.1	41.3	49	54.4	59.5
BJ007	1326.7	2380.7	2896.3	3272.4	3640.6
BK004	346.9	628.4	724.7	781.9	829
BL001	35	83.5	79.6	92.4	105.8
BL002	26.3	56.5	76.6	94.2	114.3
BL003	59.6	110.3	133.9	150.6	166.5
BO001	727.8	1343.3	1648.8	1873.5	2095
BO002	122.2	210.3	252.9	283.8	313.9
BO003	88.7	142	168.4	187.8	206.9
BP001	209.9	392.8	519.7	633.5	766.4
BQ001	157.1	274.8	337.4	385.3	434.2
BR001	29.9	46.9	53	56.7	59.8
BS001	36.6	80.9	71.1	77.8	83.9
BU002	68.4	130.9	163.4	187.9	212.6
BU003	28.2	50.7	61.1	68.5	75.6
BU004	9	16.1	19.7	22.5	25.2
BV005	7.3	17.1	24.1	30.6	38.2
BV006	43.5	73.5	88.7	100	111.3
BV007	55	98.2	123	142.8	163.9

Table 2. Attributes used in the analysis

Station	LAT (deg)	LONG (deg)	A (Km^2)	Q bar (m^3/s)	(Q bar)/A ($m^3/s/Km^2$)	Cv
AD002	46.98	68.50	14700	988.35	0.07	0.76
AD003	47.21	68.96	1360	109.75	0.08	0.79
AD004	47.36	68.33	15600	1520.45	0.1	0.78
AF002	47.04	67.74	21900	1540.48	0.07	0.79
AF003	47.34	68.14	1150	94.32	0.08	0.82
AG002	46.83	67.74	199	10.2	0.05	1.04
AH005	46.25	67.14	230	15.48	0.07	1.02
AJ001	46.47	67.59	34200	2240.85	0.07	0.81
AJ003	46.22	67.73	1210	83.72	0.07	1.1
AJ004	46.44	67.74	484	44.28	0.09	0.9
AJ010	46.34	67.47	350	26.77	0.08	0.84
AJ011	46.34	67.47	156	12.71	0.08	1.23
AK001	45.95	67.32	234	17.41	0.07	0.89
AK002	45.97	67.24	38800	1959.84	0.05	0.71
AK004	45.98	67.83	39900	2315.08	0.06	0.79
AK005	46.04	56.7	26.9	2.23	0.08	1.03
AK007	46.05	67.24	240	18.1	0.08	0.96
AK008	45.94	67.55	531	20.81	0.04	0.9
AL002	48.13	66.61	1450	119.62	0.08	0.99
AL003	46.3	67.04	8.48	0.9	0.14	1.4
AL004	46.28	67.02	3.89	0.47	0.12	1.24
AM001	45.67	66.68	557	45.41	0.08	1.24
AN001	46.3	65.71	34.4	3.05	0.09	0.68
AN002	46.29	65.72	1050	34.54	0.03	0.83
AP002	45.07	65.37	668	48.34	0.07	0.85
AP004	45.7	65.6	1100	71.27	0.06	1.11
AP006	45.5	66.32	293	29.13	0.1	0.96
AQ001	45.17	66.47	239	24.5	0.1	1.55
AR006	45.21	67.26	115	9.74	0.08	0.83
AR008	45.19	67	43	3.66	0.09	1.5
BC001	47.67	67.48	3160	256.63	0.08	0.84
BE001	47.83	66.86	2270	167.59	0.07	0.86
BJ001	47.66	65.69	363	2135.76	5.88	0.74
BJ003	47.9	66.03	510	29.61	0.06	0.82
BJ004	48.01	66.44	88.6	9.61	0.11	0.93
BJ007	47.91	86.95	7740	585.58	0.07	0.96
BK004	47.49	65.68	2090	170.53	0.06	0.81
BL001	47.65	65.58	175	14.08	0.06	1.09
BL002	47.71	65.16	173	13.7	0.06	1.29
BL003	47.44	65.11	383	27.77	0.07	0.95
BO001	46.74	65.83	5050	329.6	0.07	0.99
BO002	46.82	66.11	611	47.6	0.06	0.97
BO003	46.89	65.6	484	26.63	0.06	0.99
BP001	46.94	85.91	1340	82.18	0.06	1.36
BQ001	47.09	85.84	948	60.97	0.06	1.04
BR001	46.74	65.2	177	11.98	0.07	0.82
BS001	46.44	65.07	166	14.84	0.09	0.89
BU002	45.94	65.17	391	32.19	0.08	1.02
BU003	45.96	64.88	129	12.85	0.1	0.95
BU004	45.89	64.52	34.2	3.75	0.11	1.01
BV005	45.37	65.81	29.3	3.81	0.13	1.44
BV006	45.56	65.02	130	15.83	0.12	1.01
BV007	45.61	64.98	181	20.12	0.11	1.11

between homogeneity of the regions and the number of stations in these regions. After selection and classification of distances of each station relative to the others, the lower and upper threshold values are determined to be 0.48 and 0.89, the 30th and the 70th percentiles, respectively. Figure 2 is an example showing the position of stations forming the region of influence of station 01AM001.

After defining the regions of influence I_i for each 'target' station i, the coefficient of skewness is calculated for each station $j \in I_i$, using eqs. (16) and (17). The parameter (n) of the weight function f_{ij} (eq. 15) was chosen to be 1.25, and the parameter H to be 1.01, a value corresponding to the 80th percentile. These values determine the rate of decrease of the weighting function with an increase in the distance separating the target station i from other stations of the region. As noted by Burn (1990 a), the particular values chosen for these two parameters should not unduly affect the estimation of at-site extremes, as long as the desired shape of the weighting function is maintained. The regional coefficient of skewness $(C_s^{*_i})$ is then calculated for the target station (eq. 18). Finally, a regional $k(k^{*_i})$ is determined in order to permit the calculation of the regional 100-year flow X_{Ti}^* for the given target site (eq. 19). The regional and at-site k values as well as the quantiles are given in table 3.

A regionalization of 100-year flow as a function of basin area was performed using a linear regression. This is possible since the relation between the flow (Q) and the drainage area (A) is of the form $Q = Ca^n$, where c and n are constants. A second approach to perform the regression was analyzed using logarithms of Q and A (fig. 3). A regression line is estimated from which a regional 100-year flow was obtained. The results of the two approaches are given in table 4.

To compare the results, the variability of the 100-year flow to the sites and to the regional level in the two approaches was analyzed using the jackknife method (Quenouille, 1949). This method consists of calculating the 100-year flow several times, each time removing a recording year (on all the sites of the region that have this recording year). As mentioned earlier, the analysis is performed on eight stations, representative of the province of New Brunswick (fig. 4). To distinguish the variability obtained in the two approaches and to permit comparison of the values obtained at-site, the results for each target station are entered on a box plot diagram. The results for one of the stations (01AJ011) are presented in figs. 5 and 6.

6
Discussion

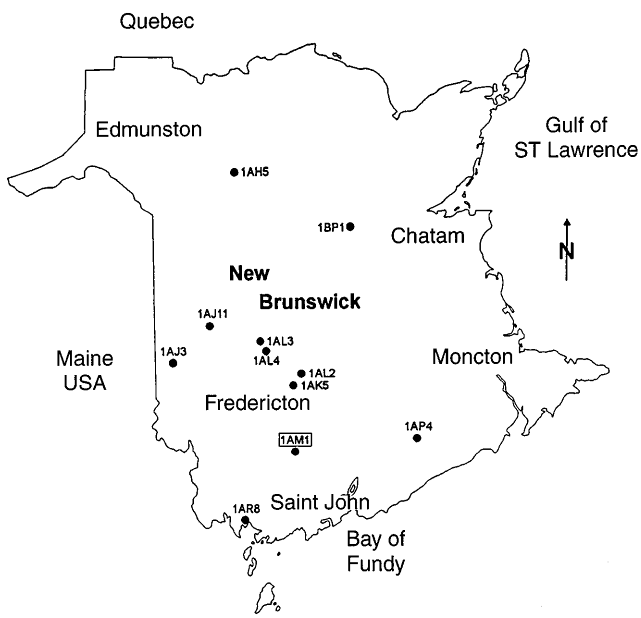
The probability distribution that describes flood flow phenomena is impossible to determine with certainty. In flood frequency analysis an attempt to choose an appropriate distribution is made, and the goodness of fit of this distribution is assessed to the data. In this study, the generalized Pareto distribution is used to adjust exceedance values by applying the method of peaks over threshold to analyze flood flow frequency in New Brunswick.

The region of influence approach is a flexible and versatile regionalization method. To obtain a degree of homogeneity within a region, the method demands careful reflection and scientific judgment. The estimation of quantile flows can be significantly improved with the appropriate selection of the threshold value S_i (eq. 10) and the attributes. An in-depth study on the optimal choice of parameters and constants may also prove useful in the application of this method.

The coefficient of skewness C_s of the generalized Pareto distribution was used first to perform the regional study. The other method used is a direct regression

Table 3. Estimated k and corresponding 100-year flood values

Station	Lambda	Alpha	Values at-site		Regional values	
			K	Flow (m^3/s)	K reg.	Flow (m^3/s)
01AD002	1.31	1316.3	0.388	4053.3	0.251	4790.7
01AD003	1.32	142.84	0.301	441	0.233	492.3
01AD004	1.42	2027.5	0.334	5569.6	0.243	6437.78
01AF002	1.43	1993.5	0.294	6482.2	0.204	7498.75
01AF003	1.06	117.02	0.242	478	0.232	483.97
01AG002	1.57	9.88	−0.034	72.3	0.064	61.50
01AH005	1.25	15.13	−0.023	102.1	0.053	88.72
01AJ001	1.27	2850.75	0.272	9918	0.196	11155.6
01AJ003	1.90	76.73	−0.084	630	−0.018	547.34
01AJ004	1.29	49.6	0.120	240.8	0.162	225.23
01AJ010	1.80	32.59	0.218	141.8	0.175	151.61
01AJ011	1.07	10.57	−0.168	97.9	−0.094	84.80
01AK001	1.82	19.61	0.127	88.7	0.142	84.16
01AK002	1.27	2000	0.484	8328.5	0.203	9063.25
01AK004	1.18	3019.9	0.304	11501	0.219	12838.74
01AK005	1.70	2.18	−0.024	15.4	0.011	14.36
01AK007	2.05	18.79	0.038	117.5	0.114	101.88
01AK008	1.00	23.36	0.122	127.3	0.156	121.75
01AL002	1.41	121.07	0.012	768.1	0.027	746.78
01AL003	2.33	0.678	−0.244	8.5	−0.265	9.07
01AL004	2.65	0.39	−0.172	4.1	−0.168	4.08
01AM001	1.46	37.38	−0.177	365.5	−0.179	368.94
01AN001	1.75	4.79	0.570	12.3	0.118	22.79
01AN002	1.40	42.26	0.223	288	0.078	334.87
01AP002	1.66	57.75	0.195	253.4	0.049	327.61
01AP004	1.67	64.73	−0.092	532	−0.022	460.39
01AP006	2.15	30.28	0.040	187.2	0.053	182.19
01AQ001	2.34	17.33	−0.293	252.6	−0.224	204.73
01AR006	2.09	12.02	0.234	46.6	0.161	52.96
01AR008	2.85	2.65	−0.277	40	−0.268	38.97
01BC001	1.42	310.27	0.209	1271.1	0.165	1363.54
01BE001	1.57	198.1	0.182	760.9	0.133	835.71
01BJ001	1.13	23.21	0.032	157.4	0.009	163.12
01BJ003	1.42	36.77	0.242	188.1	0.086	230.25
01BJ004	1.40	10.36	0.058	59.5	0.072	60
01BJ007	1.25	576.6	0.020	3640.6	0.078	3308.43
01BK004	1.12	216.11	0.267	829	0.087	1085.07
01BL001	1.83	12.92	−0.082	105.8	−0.006	90.23
01BL002	1.74	10.94	−0.201	114.3	−0.185	109.54
01BL003	1.28	29.3	0.055	166.5	0.035	172.56
01BO001	1.74	332.9	0.010	2095	0.041	1968.62
01BO002	1.39	48.91	0.028	313.9	0.050	301.96
01BO003	1.88	28.97	0.012	206.9	0.029	200.72
01BP001	1.68	63.19	−0.231	766	−0.249	802.06
01BQ001	1.41	58.62	−0.039	434.2	0.027	386.03
01BR001	2.29	14.81	0.236	59.8	0.076	80.43
01BS001	2.00	16.81	0.133	83.9	0.045	85.78
01BU002	2.89	31.48	−0.022	212.6	0.031	186.34
01BU003	2.77	13.57	0.056	75.6	0.033	79.6
01BU004	1.58	3.72	−0.011	25.23	0.020	23.79
01BV005	2.73	2.82	−0.258	38.2	−0.254	37.64
01BV006	2.67	15.76	−0.005	111.3	0.025	104.31
01BV007	3.09	18.25	−0.093	163.9	−0.030	139.66



01AM001	S
01AL004	0.1272
01AJ011	0.2920
01AL003	0.3127
01BP001	0.3621
01AK005	0.4033
01AH005	0.4391
01AP004	0.4717
01AL002	0.4722
01AJ003	0.4722
01AR008	0.4803

Fig. 2. Region of influence of station 01AM0001

of the 100-year flood on the drainage area. The jackknife sampling procedure is useful to contrast these two regionalization methods. It also allows a contrast between regional values and those estimated at-site. An examination of the variability and bias of the results is facilitated by the use of box plots. The following general observations can be drawn:

(a) The direct regression of the 100-year flood on drainage area produces results that appear to be less variable but more biased than those resulting from regionalizing C_s.
(b) The bias of the 100-year flood direct-regression method can be reduced if additional basin characteristics (other than drainage area) are used in the regression. This reduction in bias can result in increased variance. These issues need to be investigated more fully in future studies.
(c) The variance of the regionalization-of-C_s method can be reduced by a more careful application of the region of influence approach. An in-depth study of the optimal choice of parameters and constants (eqs. 10 through 15) may prove useful for variance reduction, but one should not exclude the possibility that reduction in variance can lead to increased bias.

Table 4. Comparison of 100-year flood estimates

Station	Yrs of record	Yrs removed	Sample size	100-year flood (m^3/s)		
				At-site	Regional	
					Regression	Cs Regionalization
01AD002	26–86	26	62	4059.3	4434.8	4759.7
01AD003	51–86	51	37	441	480.9	492.3
01AD004	68–79		12	5589.6	4496.3	6437.78
01AF002	30–86	30	58	6482.2	6809.5	7496.75
01AF003	62–86	80	26	476	407.4	483.97
01AG002	67–88	67	21	72.3	96.2	61.59
01AH005	72–88	72	16	102.1	116.7	88.72
01AJ001	51–86	51	37	9918	9801.5	11155.6
01AJ003	67–88	67	21	630	537.1	547.34
01AJ004	67–88	67	21	240.8	203.4	225.23
01AJ010	73–88	73	15	141.8	155.2	151.61
01AJ011	73–88	73	15	97.9	86.4	84.89
01AK001	18–41, 42–56	18, 42, 43	68	86.7	110.3	84.16
01AK002	18–67	18	49	8326.5	11766.5	9063.25
01AK004	66–88	66	22	11501	11615.0	12638.74
01AK005	65–88	65	23	15.4	16.9	14.36
01AK007	67–88	67	21	117.5	113.9	101.88
01AK008	74–88		15	127.3	235.5	121.75
01AL002	62–88		27	768.1	589.8	746.78
01AL003	70–88	70	18	8.5	7.5	9.07
01AL004	72–88		17	4.1	3.7	4.06
01AM001	62–88	62	26	365.5	305.4	366.94
01AN001	72–88	82	16	12.3	24.3	22.79
01AN002	74–88		15	288	461.1	334.87
01AP002	25–40, 61–88	25, 61, 62	41	253.4	302.8	327.61
01AP004	61–88	61	27	532	509.8	460.39

Table 4. (Contd.)

01AP006	76–88		13	187.2	133.1	182.19
01AQ001	16–88	16, 17	71	252.6	163.0	204.73
01AR006	66–88	66	22	46.6	58.6	52.96
01AR008	66–79	66	13	40	39.0	38.97
01BC001	62–88	62	26	1271.1	1132.8	1363.54
01BE001	18–88	18, 43	69	760.9	846.2	835.71
01BJ001	22–88	22, 51	65	157.4	191.7	163.12
01BJ003	64–88	64	24	186.1	226.4	230.25
01BJ004	67–83	67, 83	15	59.5	42.4	80
01BJ007	68–88	68	20	3640.6	2588.4	3306.43
01BK004	57–74	57	17	829	842.7	1085.07
01BL001	65–88	65	23	105.8	102.3	90.23
01BL002	69–88	69	19	114.3	121.6	109.54
01BL003	70–88	70	18	166.5	193.1	172.56
01BO001	18–38, 60–88	18, 38, 61	47	2095	1835.2	1988.62
01BO002	65–88	65	23	313.9	283.5	301.96
01BO003	73–88		16	206.9	241.3	200.72
01BP001	51–88	51	37	766	706.4	802.06
01BQ001	61–88	61	27	434.2	433.8	386.03
01BR001	30–32, 68–88	13, 65, 69	21	59.8	19.7	80.43
01BS001	64–88	64	24	83.9	88.0	85.78
01BU002	61–88	61	27	212.6	193.5	186.34
01BU003	62–88	62	26	75.6	75.9	79.6
01BU004	66–85	66	19	25.23	24.2	23.79
01BV005	60–71	60	11	38.2	27.5	37.64
01BV006	64–88	64	24	111.3	74.3	104.31
01BV007	67–79	67, 79	11	163.9	106.9	139.66

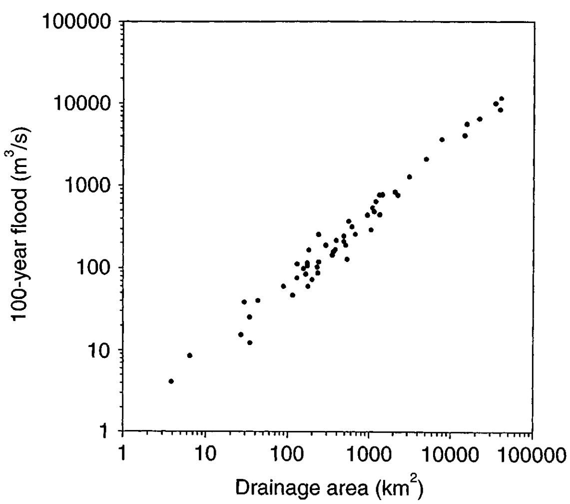


Fig. 3. Regression of the 100-year flood on the drainage area

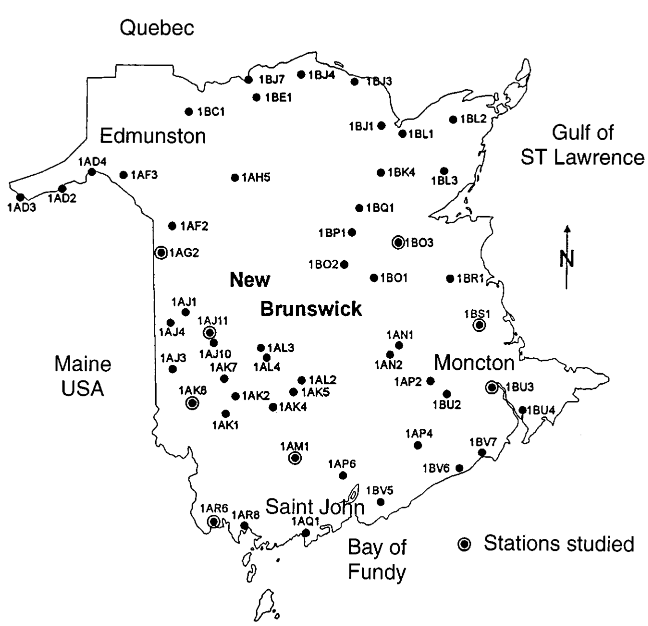


Fig. 4. Eight Stations used in the study

(d) The quantiles estimated at site (from 15 to 25 years of data) are generally less variable than those estimated regionally. This is an expected result, but the exact magnitude of this difference in variance is very much dependent on the geographical area under study, as well as on the characteristics of the hydrometric stations network.

In a regional study, a quantity to regionalize, other than the coefficient of skewness, can be the coefficient of variation, also dimensionless. It is equally preferable to take a regional value of the scale parameter α rather than the at site value to avoid a deviation between the quantiles estimated regionally and at-site.

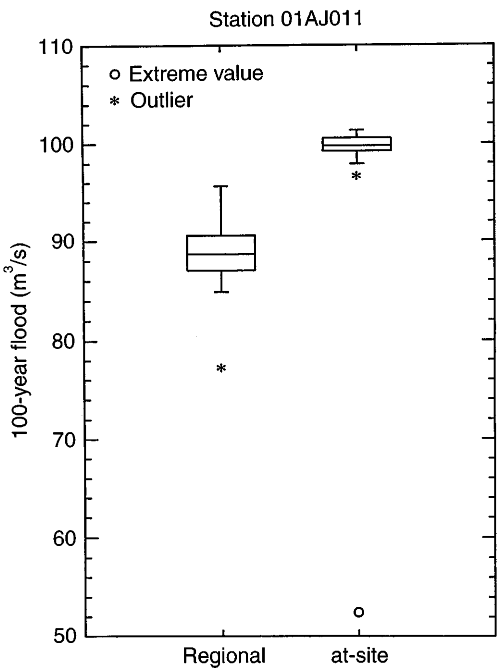


Fig. 5. Results from the regionalization of C_s

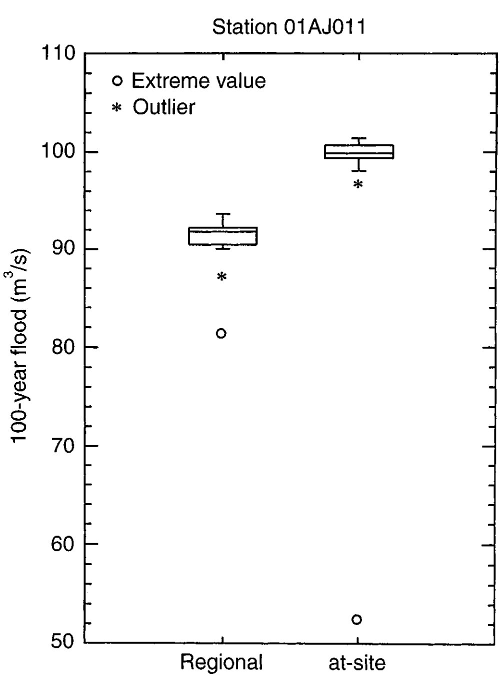


Fig. 6. Results of the regionalization by regression of the 100-year flood on drainage area

References

Acreman MC, Sinclair CD (1986) Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. Journal of Hydrology, 84, 365–380

Ashkar F, El-Jabi N, Sarraf S (1991) Study of hydrological phenomena by extreme value theory. Natural Hazards 4, 373–388

Burn DH (1989) Cluster analysis as applied to regional flood frequency. Journal of Water Resources Planning and Management, 115(5), 567–582

Burn DH (1990a) An appraisal of the "region of influence" approach to flood frequency analysis. Hydrological Sciences Journal, 35(2), 149–165

Burn DH (1990b) Evaluation of regional flood frequency analysis with a region of influence approach. Water Resources Research, 26(10), 2257–2265

Dalrymple T (1960) Flood Frequency Analysis. Manual of Hydrology: Part 3, Flood-Flow Tech., U.S. Geol. Surv. Water Supply Paper 1543 A, 11–51

Hosking J, Wallis JR (1987) Parameter and quantile estimation of the generalized Pareto distribution, Technometrics, August, 29(2), 339–349

Quenouille M (1949) Approximate tests of correlations in time series. J.R. Statistical Society, Series B, 11, 18–84

Rasmussen P, Ashkar F, Rosbjerg D, Bobée B (1994) The POT. method for flood estimation: A review. In K.W. Hipel (ed.), Stochastic and Statistical Methods in Hydrology and Environmental Engineering, Kluwer Academic Publishers, 15–26

Rosbjerg D, Madsen H, Rasmussen PF (1992) Prediction in partial duration series with generalized Pareto-distributed exceedances. Water Resour. Res., 28(11), 3001–3010

Rousselle J (1972) On some problems of flood analysis. Ph.D. Thesis, 226 pp., Colorado State Univ., Fort Collins, 1972

Simmers (1984) A systematic problem-oriented approach to regionalization. Journal of Hydrology, 73, 71–87

Tasker GD (1982) Comparing methods of hydrologic regionalization. Water Resources Bulletin, 18(6), 965–970

Todorovic P (1970) On some problems involving random number of random variables. Ann. Math. Statist., 41, 1059–1063

Wiltshire SE (1986a) Identification of homogeneous regions for flood frequency analysis. Journal of Hydrology, 84, 287–302

Wiltshire SE (1986b) Regional flood frequency analysis 1: homogeneity statistics. Hydrological Sciences Journal, 31(3), 321–333

Wiltshire SE (1986c) Regional flood frequency analysis 2: multivariate classification of drainage basins in Britain. Hydrological Sciences Journal, 31(3), 335–346

Zelenhasic E (1970) Theoretical probability distributions for flood peaks, Colorado State Univ., Hydrol. Paper no 42, 35p