Multivariate homogeneity testing in a northern case study in the province of Quebec, Canada

Fateh Chebana[*1]; Taha B.M.J.Ouarda[1]; Laura Fagherazzi[2]; Pierre Bruneau[3]; Marc Barbet[3]; Salaheddine El Adlouni [1] and Marco Latraverse[2]

[1]*Canada Research Chair on the Estimation of Hydrometeorological Variables, Hydro-Québec/NSERC Chair in statistical Hydrology, INRS-ETE, University of Quebec, 490 rue de la Couronne, Quebec, QC, Canada*

[2]*Hydro-Québec, Complexe Desjardin, Montreal, QC., Canada*

[3]*Hydro-Québec, 855 Ste-Catherine est, Montreal, QC, Canada*

[*]**Corresponding author:** Tel: (418) 654-2542

Fax: (418) 654-2600

Email: fateh_chebana@ete.inrs.ca

Abstract:

In regional frequency analysis, the examination of the regional homogeneity represents an important step of the procedure. Flood events possess multivariate characteristics which can not be handled by classical univariate regional procedures. For instance, classical procedures do not allow to assess regional homogeneity while taking into consideration flood peak, volume and duration. Chebana and Ouarda (2007) proposed multivariate discordancy and homogeneity tests. They carried out a simulation study to evaluate the performance of these tests. In the present paper, practical aspects are investigated jointly on flood peak and flood volume of a data set from the Côte-Nord region in the province of Quebec, Canada. It is shown that, after removing the discordant sites, the remaining ones constitute a homogeneous region for the volumes and heterogeneous region for the peaks. However, if both variables are jointly considered, the obtained region is possibly homogeneous. Furthermore, the results demonstrate the usefulness of the bivariate test to take into account the dependence structure between the variables representing the event, and to take advantage of more information from the hydrograph.

1. Introduction

Most hydrological events are described by several correlated variables. Multivariate representations of hydrologic phenomena include, for instance, storm duration and intensity (Yue, 2001a; Salvadori and De Michele, 2004); flood peak, volume and duration (Ashkar, 1980; Yue et al., 1999; Ouarda et al., 2000; Yue, 2001b; Shiau, 2003; De Michele et al., 2005; Zhang and Singh 2006) and drought volume, duration and magnitude (Kim et al., 2003; Ashkar et al., 1998). It is essential to understand the multivariate characteristics of such events for several engineering planning, design and management activities. Snyder (1962) and Wong (1963) can be considered as the first authors to carry out multivariate analysis in hydrology.

The study of the joint probabilistic behaviour of two or more correlated random variables that characterize the event is necessary for a thorough understanding of multivariate hydrological events. Copulas have recently been shown to represent a useful statistical tool for hydrological applications bringing the dependence behaviour between variables (e.g. Salvadori and De Michele, 2004). To represent the joint probability distribution of flood peak and volume and the joint probability distribution of flood volume and duration, Yue et al. (1999) used the Gumbel mixed model with standard Gumbel marginal distributions. Yue (2001b) and Shiau (2003) used the Gumbel logistic model with standard Gumbel marginal distributions to model flood volume and peak for different basins. El Adlouni et al. (2004) presented several copulas to model flood peak and volume with respectively Gumbel and Gamma marginal distributions.

Generally, extreme events are rare and the records are short. Consequently, the at-site frequency estimation is difficult and/or not reliable. Regional frequency analysis (RFA) is proposed as a method to overcome this lack of data. Hence, RFA is commonly used for the estimation of extreme hydrological events at sites where little or no data is available. It is based on the transfer of data available from other stations in the same hydrologic region. The delineation of hydrological homogeneous regions and the regional estimation are the two main steps of a RFA. Several authors investigated this procedure with different approaches including

Stedinger and Tasker (1986), Burn (1990), Hosking and Wallis (1993), Durrans and Tomic (1996), Nguyen and Pendey (1996), Alila (1999, 2000) and Ouarda et al., (2001). An intercomparison of various regional flood estimation procedures was presented by GREHYS (1996a,b).

The literature on flood frequency analysis can be classified into four classes according to the local/regional and univariate/multivarite aspects. In the first two classes there are the local-univariate and regional-univariate studies where many references can be found in the literature. The third class contains local-multivariate flood frequency analysis (e.g., Ashkar, 1980; Yue et al., 1999; Ouarda et al., 2000; Yue, 2001b; Shiau, 2003; De Michele et al., 2005 and Zhang and Singh, 2006). However, very little attention has been given to the forth class which consists in regional-multivariate studies (Ouarda et al., 2000 and Chebana and Ouarda, 2007). Ouarda et al. (2000) carried out a canonical correlation analysis procedure for a joint regional study of flood peak and volume in the province of Quebec, Canada.

Chebana and Ouarda (2007) proposed discordancy and homogeneity tests based on multivariate *L*-moments and copulas. The proposed multivariate discordancy and homogeneity tests are extensions of those given by Hosking and Wallis (1993). Chebana and Ouarda (2007) evaluated the performances of these multivariate tests using simulations. They demonstrated, for a given extreme event, the importance of jointly and simultaneously treating all variables and identifying a single homogeneous region. In the present paper, practical aspects of these multivariate tests are studied based on a real-word regional data set. The data set corresponds to sites from the Côte-Nord region in the eastern part of the province of Quebec, Canada. The application is carried out on flood event and the focus is on the volume and peak variables.

The paper is organized as follows. Section 2 contains the theoretical background, including flood characteristics, bivariate modeling, multivariate *L*-moments and the discordancy and homogeneity tests. Section 3 is devoted to the description of the case study. The procedure

followed in this study is presented in Section 4 whereas Section 5 deals with the corresponding results. Concluding remarks are presented in Section 6.

2. Background

In this section, the background elements to apply the multivariate discordancy and homogeneity tests are presented. Flood characteristics, bivariate modeling including copulas and marginal distributions, and the discordancy and homogeneity tests are briefly described.

2.1 Flood characteristics

In Figure 1, a typical flood hydrograph is illustrated. A flood hydrograph is mainly characterized by its volume, duration and peak. Flood duration has to be determined first in order to compute flood volume. Flood duration can be determined whenever the start date s_i and end date e_i are identified for the ith series as $D_i = e_i - s_i$. The annual flood volume series can be constructed using the following formula (see e.g., Yue et al., 1999):

$$V_i = \sum_{j=s_i}^{e_i} q_{ij} - \frac{1}{2}(q_{is} + q_{ie}),\ i = 1,2,..., \qquad (1)$$

where q_{ij} represents the observed streamflow value at the jth day of the ith year, q_{is} and q_{ie} are respectively the observed daily streamflow values on the start date and end date of flood runoff for the ith year. The annual flood peak series is given by

$$Q_i = \max\left\{q_{ij},\ j = s_i, s_i + 1,..., e_i\right\} \qquad (2)$$

2.2 Bivariate flood modeling

In bivariate modeling, one should obtain a joint bivariate distribution for the variables. However, one should also distinguish the dependence structure from the margins. To this end, one needs to specify three elements: a copula to describe the dependence structure between the two random variables, along with a marginal distribution for each variable.

2.2.1 Copulas

In the remainder of the paper we denote F_1 and F_2 the marginal distribution functions of given random variables $X^{(1)}$ and $X^{(2)}$, and $F_{1,2}$ is the joint distribution function of $\left(X^{(1)}, X^{(2)}\right)$.

Independently of the marginal distributions, a copula is a description and a model of the dependence structure between the two random variables. To overcome the limitations of classical dependence measures, copulas have recently received increasing attention in various science fields (see for instance Nelsen, 1999). A copula is a function C: $I \times I \to I$ (I = [0, 1]) such that:

- for all $u, v \in I$: $C(u, 0) = 0$, $C(u, 1) = u$, $C(0, v) = 0$, and $C(1,v) = v$;
- for all $u_1, u_2, v_1, v_2 \in I$ such that $u_1 \le u_2$ and $v_1 \le v_2$: $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \ge 0$

The link between copulas and bivariate distributions is provided by Sklar's (1959) result. It states that there exists a copula C such that:

$$F_{1,2}(x_1, x_2) = C\left(F_1(x_1),\ F_2(x_2)\right) \quad \text{for all real } x_1 \text{ and } x_2 \tag{3}$$

When F_1 and F_2 are continuous, the copula C is unique.

Archimedean and extreme value (EV) copulas represent classes of particular interest. The class of EV copulas is given by the formula (Pickands, 1981):

$$C(u,v) = \exp\left\{(\log u + \log v)\, A\left(\frac{\log u}{\log u + \log v}\right)\right\}, \quad 0 < u, v < 1 \tag{4}$$

where the dependence function A is convex and defined on [0, 1] with $\max\{t, 1-t\} \le A(t) \le 1$. A bivariate Archimedean copula is characterized by the expression:

$$C(u,v) = \psi^{-1}\left(\psi(u) + \psi(v)\right), \quad 0 < u, v < 1 \tag{5}$$

where the generator $\psi(.)$ is a convex decreasing function satisfying $\psi(1) = 0$.

As it is already shown in previous studies, e.g. Salvadori and De Michele (2004), Archimedean copulas represent convenient multivariate models for hydrological flood events.

When the multivariate context is involved, some practical questions can be raised regarding copulas, for instance:

- How a copula can be fitted to a given sample?
- How copula's parameters can be estimated?
- And how a sample can be generated from a model defined through a given copula?

Partial answers to these questions are given for the Archimedean and extreme value copulas.

First, the fitting problem is resolved for Archimedean copulas. According to Genest and Rivest (1993), an Archimedean copula, with a generator function ψ, is characterized by the following function:

$$K_{\psi}(z) = z - \frac{\psi(z)}{\psi'(z)} \tag{6}$$

which can be estimated by:

$$\widehat{K}(z) = \frac{1}{N}\sum_{i=1}^{N} 1_{[w_i \le z]} \quad \text{where} \quad w_i = \frac{1}{N-1}\sum_{t=1}^{N} 1_{[x_1^t < x_1^i, x_2^t < x_2^i]},\ i = 1,..,N \tag{7}$$

for a given bivariate sample $(x_1^1, x_2^1),(x_1^2, x_2^2),...,(x_1^N, x_2^N)$. It is shown in Genest and Rivest (1993) that $\widehat{K}$ is a consistent estimator of *K* under weak regularity conditions.

It is shown in several studies (e.g. Yue, 2001b and Shiau, 2003) that an interesting copula to model flood characteristics is the Gumbel logistic copula given by:

$$C_m(x, y) = \exp\left\{-\left[(-\log x)^m + (-\log y)^m\right]^{1/m}\right\},\ m \ge 1,\ 0 \le x, y \le 1 \tag{8}$$

which is an Archimedean copula with generator function $\psi(x) = (-\log x)^m$, and it is also an extreme value copula with dependence function $A(t) = \left(t^m + (1-t)^m\right)^{1/m}$. The corresponding function *K* defined in (6) for the Gumbel logistic copula is given by $K_m(z) = z - \frac{z \log(z)}{m}$.

Second, the parameter estimation problem for Archimedean copulas is also resolved. In particular, the parameter *m* of the copula C_m is related to the correlation coefficient ρ through the equation (Gumbel and Mustafi, 1967):

$$m = \frac{1}{\sqrt{1-\rho}}, \quad 0 \le \rho < 1 \tag{9}$$

Hence, it can be estimated by a plug-in of the empirical version of the correlation coefficient in equation (9). However, it can also be estimated by:

$$\hat{m} = 1 + \frac{\hat{\tau}_{1,2}}{1-\hat{\tau}_{1,2}} \tag{10}$$

where $\hat{\tau}_{1,2}$ is the empirical estimator of

$$\tau_{1,2} = 4E\left[F_{1,2}(X^{(1)}, X^{(2)})\right] - 1 \tag{11}$$

which is a version of the Kandall's tau coefficient for the random vector $\left(X^{(1)}, X^{(2)}\right)$. A simple estimator of the Kandall's tau coefficient is given by $\hat{\tau}_{1,2} = 4\bar{G} - 1$ where $\bar{G}$ is the mean of the "pseudo-sample" $G_i = \frac{1}{n-1}\#\left\{(X_j^{(1)}, X_j^{(2)}) : X_j^{(1)} < X_i^{(1)}, X_j^{(2)} < X_i^{(2)}\right\}, \quad i = 1, ..., n$ (see Genest and Rivest, 1993).

Regarding the last question, related to the generation of samples from the variables $\left(X^{(1)}, X^{(2)}\right)$ according to the extreme value copula, an algorithm is developed by Ghoudi et al. (1998). The algorithm is summarized in the following. Let U_1, U_2 be uniform random variables and Z be a random variable with a cumulative distribution function G_Z and probability density function g_Z given by $G_Z(z) = z + z(1-z)A'(z)/A(z), 0 \le z \le 1.$ This algorithm consists of the following steps:

1. Simulate *Z*;
2. Given *Z*, take $W = U_1$ with probability $p(Z)$ and $W = U_1 U_2$ with probability $1 - p(Z)$, where $p(z) = z(1-z)A''(z)/\left(A(z)g_Z(z)\right)$;
3. Set $X^{(1)} = W^{Z/A(Z)}$ and $X^{(2)} = W^{(1-Z)/A(Z)}$.

When using this algorithm in practice it is important to take into consideration the numerical nonparametric smoothing, since it depends on functions related to the first and second derivatives of the function *A*. Despite the general validity of this procedure, extra information about the model, e.g. parametric form of *A*, can be useful to increase the speed and accuracy of the generation algorithm.

2.2.2 Marginal modeling

The 2-parameter Gumbel distribution can be used to model the marginal flood variables (Yue, 2001b and Shiau, 2003). However, as it is indicated in Hosking and Wallis (1993) and Chebana and Ouarda (2007), it is preferable to employ a 4-parameter Kappa distribution for the homogeneity test. Its cumulative distribution function is given by:

$$F(x) = \left[1 - h\left(1 - \kappa \frac{(x-u)}{\alpha} \right)^{\frac{1}{\kappa}} \right]^{\frac{1}{h}} \tag{12}$$

with parameters *u* (position), α (scale), κ and *h* (shape).

The parameters of the Kappa distribution can be estimated by the *L*-moment method (Hosking and Wallis, 1997). Indeed, if we denote respectively by λ_k the *L*-moment and t_k the *L*-moment coefficient of order *k*, the first Kappa *L*-moments are given by:

$$\begin{aligned} \lambda_1 &= u + \alpha(1 - g_1)/\kappa \\ \lambda_2 &= \alpha(g_1 - g_2)/\kappa \\ t_3 &= (-g_1 + 3g_2 - 2g_3)/(g_1 - g_2) \\ t_4 &= (-g_1 + 6g_2 - 10g_3 + 5g_4)/(g_1 - g_2), \end{aligned} \tag{13}$$

where

$$g_r = \begin{cases} \dfrac{r\Gamma(1+\kappa)\Gamma(r/h)}{h^{1+\kappa}\Gamma(1+\kappa+r/h)}, & h > 0 \\ \dfrac{r\Gamma(1+\kappa)\Gamma(-\kappa-r/h)}{(-h)^{1+\kappa}\Gamma(1-r/h)}, & h < 0 \end{cases} \tag{14}$$

Therefore its parameters can be estimated by the use of equations (13) but there are no simple and direct expressions. Hosking (1996) developed a routine to find numerically the Kappa *L*-moment parameter estimators.

2.3 Multivariate *L*-moments

Instead of traditional moments, for statistical inference of hydrological variables, the *L*-moment approach offers strong advantages for modeling heavy-tailed distributions. For a review related to *L*-moments the reader can consult Hosking and Wallis (1997). Multivariate *L*-moments are principally developed by Serfling and Xiao (2007).

By analogy with the covariance representation of the *L*-moment of order k, multivariate *L*-moments are matrices Λ_k with *L*-comoment elements defined by:

$$\lambda_{k[ij]} = \mathrm{Cov}\left(X^{(i)}, P^*_{k-1}\left(F_j(X^{(j)})\right)\right),\ i,j = 1,2 \text{ and } k = 2,3,\ldots \tag{15}$$

where P^*_k is the so-called shifted Legendre polynomial. As it can be seen, the elements $\lambda_{k[ij]}$ and $\lambda_{k[ji]}$ are not necessarily equal. The first *L*-comoment elements are given by:

$$\begin{aligned}
\lambda_{2[12]} &= 2\mathrm{Cov}\left(X^{(1)}, F_2(X^{(2)})\right) \\
\lambda_{3[12]} &= 6\mathrm{Cov}\left(X^{(1)}, \left(F_2(X^{(2)}) - 1/2\right)^2\right) \\
\lambda_{4[12]} &= \mathrm{Cov}\left(X^{(1)}, 20\left(F_2(X^{(2)}) - 1/2\right)^3 - 3\left(F_2(X^{(2)}) - 1/2\right) + 1\right)
\end{aligned} \tag{16}$$

The *L*-comoment coefficients are given by:

$$\tau_{2[12]} = \frac{\lambda_{2[12]}}{\lambda_1^{(1)}} \quad \text{and} \quad \tau_{k[12]} = \frac{\lambda_{k[12]}}{\lambda_2^{(1)}},\ \text{for } k = 3,4,\ldots \tag{17}$$

where $\lambda_k^{(j)} = \lambda_{k[jj]}$ is the classical univariate *k*th *L*-moment of the variable $X^{(j)}$. The matrix of the *L*-comoment coefficients is written as:

$$\Lambda_k^* = \left(\tau_{k[ij]}\right)_{i,j=1,2} = \begin{pmatrix} \tau_{k[11]} & \tau_{k[12]} \\ \tau_{k[21]} & \tau_{k[22]} \end{pmatrix},\ \text{for } k = 2,\ 3,\ldots \tag{18}$$

and for k = 1, the first order bivariate L-moment corresponds to the mean vector $\lambda_1 = E(X^{(1)}, X^{(2)})^t$.

2.4 Discordancy and homogeneity tests

2.4.1 Discordancy

A preliminary screening step, before proceeding with the homogeneity analysis, consists in identifying discordant sites among a set of N sites. A multivariate extension of the Hosking and Wallis (1993) discordancy test is proposed by Chebana and Ouarda (2007). It is defined for each site i using the matrix $U_i^t = \left[\Lambda_2^{*(i)} \ \Lambda_3^{*(i)} \ \Lambda_4^{*(i)} \right]$ which is composed by the three L-moment matrices $\Lambda_2^{*(i)}$, $\Lambda_3^{*(i)}$ and $\Lambda_4^{*(i)}$ defined by (18). Hence, a site i is discordant, with respect to the considered set of sites, if $\|D_i\|$ takes large values, where:

$$D_i = \frac{1}{3}\left(U_i - \overline{U}\right)^t S^{-1}\left(U_i - \overline{U}\right), \tag{19}$$

$$S = \frac{1}{N-1}\sum_{i=1}^{N}\left(U_i - \overline{U}\right)\left(U_i - \overline{U}\right)^t, \tag{20}$$

$$\overline{U} = \frac{1}{N}\sum_{i=1}^{N} U_i, \tag{21}$$

$\|A\|$ denotes the spectral norm of a matrix A given by $\|A\| = \sqrt{\text{maximum eigenvalue of } A^t A}$ and A^t is the transpose of a matrix or a vector A. Note that Chebana and Ouarda (2007) considered other matrix norms and indicated that no significant difference was observed in the results obtained with the other norms.

The constant $c = \chi_{1-0.05}(3)/3 = 2.6$ may be considered as a critical value for $\|D_i\|$ for large regions, where $\chi_{1-\alpha}(d)$ denotes the quantile of a chi-square distribution of order α with d degrees of freedom. Chebana and Ouarda (2007) proposed the use of a bootstrap technique to determine a critical value for short values of N. Hosking and Wallis (1997) advised to examine the data for sites with the largest $\|D_i\|$ values, regardless of the magnitude of these values.

2.4.2 Homogeneity test

The following multivariate homogeneity test is proposed by Chebana and Ouarda (2007). It is an extension of the univariate test proposed by Hosking and Wallis (1993). It can be summarized in the followings. Let $V_{\|.\|}$ be the statistic defined as:

$$V_{\|.\|}{}^2 = \frac{\sum_{i=1}^{N} n_i \left\| \Lambda_2^{*(i)} - \overline{\Lambda_2^*} \right\|^2}{\sum_{i=1}^{N} n_i} \tag{22}$$

where $\overline{\Lambda_2^*} = \left(\sum_{i=1}^{N} n_i \right)^{-1} \sum_{i=1}^{N} n_i \Lambda_2^{*(i)}$ and $\Lambda_2^{*(i)}$ is the *L*-covariation coefficient matrix for site *i*, with record length n_i, $i = 1,...,N$. In order to get interpretable results of the computed value of the statistic $V_{\|.\|}$ from the observations, it is convenient to standardize it by the use of a large number of simulated homogeneous regions. The simulated regions are homogeneous with sites having the same record lengths as their observed counterparts. Hence, the statistic that measures the heterogeneity of a set of sites is given by:

$$H_{\|.\|} = \frac{V_{\|.\|} - \mu_{Vsim}}{\sigma_{Vsim}} \tag{23}$$

where μ_{Vsim} and σ_{Vsim} are respectively the mean and standard deviation of the N_{sim} values of $V_{\|.\|}$ of simulated regions. The EV or Archimedean copulas with the marginal 4-parameter Kappa distributions are the bivariate distributions on which the simulations are carried out to compute μ_{Vsim} and σ_{Vsim}. A region of sites is declared to be homogeneous if $H_{\|.\|} < 1$, acceptably homogeneous if $1 < H_{\|.\|} < 2$ and definitely heterogeneous if $H_{\|.\|} > 2$. Note that in the univariate framework, the statistics $V_{\|.\|}$ and $H_{\|.\|}$ are equivalent to the classical statistics defined by Hosking and Wallis (1993). For more details concerning the multivariate homogeneity test, the reader is referred to Chebana and Ouarda (2007).

The test statistic $H_{\|.\|}$ is standardized on the basis of the mean and standard-error of N_{sim} simulated homogeneous regions. The value of N_{sim} = 500 is shown to be appropriate to allow the test to perform well. However, higher values of N_{sim} allow to improve the estimation of μ_{Vsim} and σ_{Vsim} and hence to make the right decisions when the values of $H_{\|.\|}$ are close to the thresholds 1 and 2.

3. Case study

The application of the multivariate discordancy and homogeneity tests concerns a regional data set of interest for the Hydro-Québec company. The phenomenon to be studied is the flood, with bivariate characteristics, that is, volume *V* and spring peak *Q*. The data is from sites of the Côte Nord in the north part of the province of Quebec, Canada. The data set counts *N* = 26 stations with record lengths n_i from 14 to 48. Some information about the data are given in Table 1. The geographical location of the underlying sites is presented in Figure 2.

4. Study procedure

The procedure of the study is composed in the following three main steps:

1. *Correlation:* Assessment of the correlation coefficient between the variables *V* and *Q* for each site.
2. *Discordancy:*
    a. Evaluation of the discordancy of each site on the basis of each variable *V* and *Q* and on the joint variable (*V*,*Q*).
    b. Identification of the corresponding discordant sites.
3. *Homogeneity:* To carry out the homogeneity test given in equation (23), the following steps are required:
    a. Removing the discordant sites identified in 2.b, and using the remaining sites in the following steps.

b. Modeling the joint variable (*V*,*Q*): Employing the Archimedean copula characterization (7). The marginal of both *V* and *Q* is taken to be the Kappa distribution (12).

c. Estimating the model parameters: For the marginal Kappa distribution the estimation is based on the *L*-moment ratios $1, \overline{t}_2, \overline{t}_3, \overline{t}_4$ with $\overline{t}_k = \sum_i n_i t_k^{(i)} \Big/ \sum_i n_i$. Their parameters are estimated using equations (13) and (14). The parameter *m* of the copula (8) can be estimated by:

$$\hat{m} = 1 + \frac{\overline{\tau}_{1,2}}{1-\overline{\tau}_{1,2}}, \text{ where } \overline{\tau}_{1,2} = \frac{\sum_{i=1}^{N} n_i \hat{\tau}_{1,2}^{(i)}}{\sum_{i=1}^{N} n_i} \qquad (24)$$

where $\hat{\tau}_{1,2}^{(i)}$ is the empirical at-site estimator of $\tau_{1,2}$ defined in (11).

d. Computation of the homogeneity tests :

i. Computation of observed $V_{\| \|}$ statistics (equation 22) both on the margins and on the joint variables.

ii. Computation of the statistics V_{sim}, based on N_{sim} = 500 generated homogeneous regions. The bivariate regions are generated according to the model defined through the Gumbel logistic copula (8) and Kappa distribution (12) for the margins. For both *V* and *Q* univariate regions are generated according to the inversion formula of (12). The sampling procedure for the bivariate samples is based on the algorithm developed by Ghoudi et al. (1998) and the inversion formula of (12).

iii. Assessment of the mean μ_{sim} and the standard deviation σ_{sim} on the basis of the N_{sim} values of V_{sim}.

iv. Evaluation of homogeneity statistics $H_{\| \|}$, respectively for the variables *V*, *Q* and (*V*,*Q*), by combining $V_{\| \|}$, μ_{sim} and σ_{sim} using the expression (23).

5. Results and discussion

Results of the correlation between the variables *V* and *Q* are reported in Table 1. Their values are positive and generally exceed 0.5. This means that the variables *V* and *Q* are generally highly correlated. Hence the use of bivariate analysis is of interest compared to the use of two univariate homogeneity tests. Table 2 presents the discordancy results. The sites that may be discordant are identified and their respective *L*-moment values that may cause the discordancy are also identified. Namely, sites 2 and 16 for *V*; site 2 or sites 2 and 3 for *Q*; and sites 2 and 21 for (*V, Q*). These sites are eliminated to allow application of the respective homogeneity tests. Clearly, the discordant sites are not the same for *V, Q* or (*V,Q*).

The application of the Archimedean copula characterization (6) with the estimate (7) leads to fit the Gumbel logistic copula to the bivariate data of each site. The illustration of this fitting is presented in Figure 3 for each site of the data set. Regarding the marginal distributions, the fitting is based on the empirical cumulative distribution function. One of the most used expressions is given by Cunnane (1978):

$$P_k = \frac{k - 0.4}{N + 0.2} \tag{25}$$

for a sorted sample $x_{(1)} \le ... \le x_{(k)} \le ... \le x_{(N)}$. Figure 4 illustrates, as an arbitrary selected example, the fitting of the marginal distributions and the dependence structure for the data of station 02RF001 (the 7th station). Hence, for both variables *V* and *Q* the samples are fitted with the Gumbel distribution, for which the cumulative distribution function is given by:

$$F(x) = \exp\left\{-\exp\left(-\tfrac{x-\beta}{\alpha}\right)\right\},\ x \text{ real},\ \alpha > 0 \text{ and } \beta \text{ real} \tag{26}$$

The use of the Gumbel distribution in this context was discussed in several previous studies (e.g., Yue et al., 1999; Yue, 2001b and Shiau, 2003). Note that the Gumbel distribution is covered by the Kappa distribution, since the latter contains the generalized extreme value as a special case.

The detected discordant sites present some special characteristics in terms of their *L*-moments or their physiographical attributes. Site 2 is discordant for all variables. Site 2 has the

lowest mean l_1, the largest *L*-CV t_2 and the lowest *L*-skewness t_3 for the volume. However, for the peak, site 2 has the lowest mean l_1, the largest *L*-CV t_2 and the lowest *L*-kurtosis t_4. Furthermore, site 2 represents the smallest basin in the region. Also, 50% of its basin area is controlled by a reservoir. All this could explain its rejection. Site 3, which may be discordant for *Q*, has a very small value of the *L*-skewness t_3. Site 16 is particular because of its short record length (n = 16, Table 1) as well as the small values of the three characteristics t_2, t_3 and t_4 for the volume. Site 21 is detected to be discordant for the joint variables (*V*,*Q*) because of the poor fitting of the Gumbel logistic copula as can be seen from Figure 3.

After removing the discordant sites, the weighted *L*-moments and corresponding model parameters are computed. These results are presented in Table 3. The estimated parameters are those of the distribution of the homogeneous generated regions.

Homogeneity results are reported in Table 4. The decision about the homogeneity of the remained sites is taken according to the values of the test statistics $H_{\|\cdot\|}$. In the bivariate framework, the region is possibly homogeneous since the value of $H_{\|\cdot\|}$ is in the range $]1,2[$. When considering only the volume, the value of the statistic $H_{\|\cdot\|}$ is less than 1. Hence, the region is homogeneous. As for the peak, the region is declared to be heterogeneous since the statistic $H_{\|\cdot\|}$ has a value greater than 2. Note that the decision concerning the regional homogeneity in relation to the peak variable remains the same whether site 3 is removed or not. Indeed, from Table 3, we observe that removing site 3 has no effect on the values of the weighted *L*-moments and has no significant effect on the values of Kappa parameter estimates.

The advantage of choosing a Kappa distribution is that it allows to include several distributions used in hydrology. This in turn allows to fit the same distribution to all sites avoiding hence the subjective choice of a different distribution for each site. However, a disadvantage of this choice is the high number of parameters to be estimated, especially in the

multivariate context. Consequently, the estimation of these parameters increases the uncertainty of the model. In order to reduce this uncertainty, based on the parsimony principle, one looks for a model with the smallest number of parameters.

6. Conclusions and recommendations

The multivariate discordancy and homogeneity tests based on *L*-moments are applied to a set of sites from the Côte-Nord region in the eastern part of the province of Quebec, Canada. These tests are proposed by Chebana and Ouarda (2007) where a simulation study was carried out to evaluate their performance. The main conclusion highlights the importance of considering jointly and simultaneously all variables characterizing the extreme event and hence identifying a single homogeneous region. In the present paper, practical aspects of these multivariate tests are investigated jointly on flood peak and flood volume. Some of such aspects include the selection of the bivariate distribution using goodness-of-fit tests for the each marginal distribution as well as for the copula, the estimation of the corresponding parameters and a description of the discordant sites.

After removing the discordant sites, the remaining ones represent a homogeneous region for the volume, heterogeneous region for the peak and possibly homogeneous region if both variables are simultaneously considered. The results concerning the volume present a certain level of concordance with the bivariate results with respect to the values of the statistics $H_{\|\cdot\|}$. Physically, this can be explained by the fact that the volume contains more information than the peak concerning the whole hydrograph. This shows the interest of the bivariate test to take into account the dependence structure between the variables, and to take advantage of more information from the hydrograph. The implementation and the use of the multivariate tests are simple similarly to the univariate ones. The corresponding Matlab® programs are available from the authors on request.

Acknowledgment

The authors wish to thank Hydro-Québec and the Natural Sciences and Engineering Research Council (NSERC) of Canada for the financial support of this study. The authors wish also to thank the Editor-In-Chief and three anonymous reviewers whose comments helped improve the quality of the paper.

References

Alila, Y. (1999) A hierarchical approach for the regionalization of precipitation annual maxima in Canada. *J. Geophys. Res.*, **104**, 31,645-31,655.

Alila, Y. (2000) Regional rainfall depth-duration-frequency equations for Canada. *Water Resour. Res.*, **36**, 1767-1778.

Ashkar, F. (1980) *Partial duration series models for flood analysis*. PhD thesis, Ecole Polytechnique of Montreal, Montreal, Canada.

Ashkar, F.; El Jabi, N. and Issa, M. (1998) A bivariate analysis of the volume and duration of low-flow events. *Stoch. Hydrol. Hydraul.*, **12**, 97-116.

Burn, D. H. (1990) Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resour. Res.*, ***26***, 2257-2265.

Chebana, F. and Ouarda, T.B.M.J. (2007) Multivariate *L*-moment homogeneity test. *Water Resour. Res.*, **43**, W08406, doi:10.1029/2006WR005639.

Cunnane C. (1978) Unbiased plotting positions—a review. *J. Hydrology*, **37**, 205–222.

De Michele, C.; Salvadori, G.; Canossi, M.; Petaccia, A. and Rosso, R. (2005) Bivariate Statistical Approach to Check Adequacy of Dam Spillway. *J. Hydrologic Engrg.*, **10**, 50-57.

Durrans, S.R. and Tomic, S. (1996) Regionalization of low-flow frequency estimates: an Alabama case study. *Water Resour. Bull.*, **32**, 23-37.

El Adlouni, S.; Favre, A-C. and Bobée, B. (2004) Multivariate frequency analysis using Copulas. *DeMoSTAFI Conference "Dependence Modelling Statistical Theory and Application in Finance and Insurance"*. Québec, Canada 20-22 Mai 2004.

Genest, C. and Rivest, L-P. (1993) Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association*, **88**, 1034-1043.

Ghoudi, K.; Khoudraji, A. and Rivest, L-P. (1998) Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. *Canad. J. Statist.*, **26**, 87-197.

GREHYS (1996a) Presentation and review of some methods for regional flood frequency analysis. *J. Hydrology,* **186**, 63-84.

GREHYS (1996b) Inter-comparison of regional flood frequency procedures for Canadian rivers. *J. Hydrology,* **186**, 85-103.

Gumbel, E.J. and Mustafi, C.K. (1967) Some analytical properties of bivariate extreme distributions. *J. Amer. Statist. Assoc.*, **62**, 569–588.

Hosking, J. R. M. and Wallis, J. R. (1993) Some statistics useful in regional frequency analysis. *Water Resour. Res.,* **29**, 271-282.

Hosking, J. R. M. (1996) Fortran routines for use with the method of L-moments, Version 3. IBM Research Report RC20525. (Description of the routines in this library.)

Hosking, J. R. M. and Wallis, J. R. (1997) *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press. 240 pages.

Kim, T.; Valdés, J. B. and Yoo, C. (2003) Nonparametric Approach for Estimating Return Periods of Droughts in Arid Regions. *J. Hydrologic Engrg.*, **8**, 237-246.

Nelsen, R. B. (1999) *An introduction to copulas*. Lecture Notes in Statistics, 139. Springer-Verlag, New York. 236 pages.

Nguyen, V.-T.-V. and Pandey, G. (1996) A new approach to regional estimation of floods in Quebec. In: Delisle, C.E., Bouchard, M.A. (Eds.), Proceedings of the 49th Annual Conference of the CWRA, June 26–28, Quebec City. Collection Environnement de l'U. de M., 587–596.

Ouarda, T. B. M. J.; Haché, M.; Bruneau, P. and Bobée, B. (2000) Regional Flood Peak and Volume Estimation in Northern Canadian Basin. *ASCE J. Cold. Reg. Engrg.,* **14**, 176-191.

Ouarda, T. B. M. J.; Girard, C.; Cavadias, G. S. and Bobée, B. (2001) Regional flood frequency estimation with canonical correlation analysis. *J. Hydrology*, **254,** 157-173.

Pickands, J. (1981) Multivariate extreme value distributions. In *Bulletin of the International Statistical Institute*: Proceedings of the 43rd Session (Buenos Aires), pp. 859-878. Voorburg, Netherlands: ISI.

Salvadori, G. and De Michele, C. (2004) Analytical calculation of storm volume statistics involving Pareto-like intensity-duration marginals. *Geophys. Res. Lett.*, **31**, L04502.1-L04502.4.

Serfling, R. and Xiao, P. (2007) A Contribution to Multivariate L-Moments: L-Comoment Matrices. *Journal of Multivariate Analysis*, to appear.

Shiau, J. T. (2003) Return period of bivariate distributed extreme hydrological events. *Stoch. Environ. Res. Risk. Assess.*, **17**, 42-57.

Sklar, A. (1959) Fonctions de répartition à *n* dimensions et leurs marges *Publ. Inst. Statist. Univ. Paris*, **8**, 229-231.

Snyder, W. M. (1962) Some possibilities for multivariate analysis in hydrologic studies. *J. Geophys. Res.*, **67**, 721-729.

Stedinger, J.R. and Tasker, G. (1986) Regional hydrologic analysis, 2, Model-error estimators, estimation of sigma and log Pearson type 3 distributions. *Water Resour. Res.*, **22**, 1487-1499.

Yue, S.; Ouarda, T. B. M. J.; Bobée, B.; Legendre, P. and Bruneau, P. (1999) The Gumbel mixed model for flood frequency analysis. *J. Hydrology*, **226**, 88-100.

Yue, S. (2001a) The Gumbel logistic model for representing a multivariate storm event. *Adv. Water Res.*, **24**,179-185.

Yue, S. (2001b) A Bivariate Extreme Value Distribution Applied to Flood Frequency Analysis. *Nord. hydrol.*, **32**, 49-64.

Wong, S. T. (1963) A multivariate statistical model for predicting mean annual flood in New England. *Annal. Assoc. Am. Geographer*, **53**, 298-311.

Zhang, L. and Singh, V. P. (2006) Bivariate Flood Frequency Analysis Using the Copula Method. *J. Hydrologic Engrg.*, **11**, 150-164.

List of tables

List of figures

Table 1. Information concerning the data set.

#	Station number	Station name	Area (km^2)	n_i	(V,Q) correlation coefficient
1	02RH049	Petit Saguenay	729	24	0.50
2	02RH048	Des Ha Ha	564	19	0.73
3	02RH034/35	Aux Écorces	1120	34	0.50
4	02RH027	Pikauba	489	34	0.34
5	02RG005	Métabetchouane	2270	30	0.54
6	02RC011	Petite Péribonka	1090	31	0.62
7	02RF001	Chamouchouane (Ashuapmushuan)	15300	43	0.70
8	02RD002	Mistassibi	8690	39	0.52
9	02RD003	Mistassini	9620	43	0.52
10	02RD004	Manouane	3720	23	0.39
11	02RH045	Valin	740	31	0.42
12	02RH046	Ste-Marguerite	1100	21	0.48
13	02SC001	DesEscoumins	779	19	0.49
14	02SC002	Portneuf	2580	20	0.80
15	02UA003	Godbout	1570	30	0.75
16	02UC003	Aux-Pékans	3390	16	0.54
17	02VA001/3	Tonerre	674	40	0.64
18	02VB004	Magpie	7200	27	0.66
19	02VC001	Romaine	13000	48	0.68
20	02WA001	Nabisipi	2060	25	0.78
21	02WA002	Aguanus	5590	19	0.60
22	02WB002	Natashquan	15600	39	0.75
23	02WC001	Etamamiou	2950	19	0.82
24	02XB001	St Augustin	5750	14	0.73
25	02XC001	St Paul	6630	25	0.73
26	02UC002	Moisie	19000	39	0.65

Table 2. Values of *L*-moments and discordancy statistics for each site in the region.

	Volume					Peak					Volume and Peak
# site	l_1	t_2	t_3	t_4	D_V	l_1	t_2	t_3	t_4	D_Q	$D_{V,Q}$
1	234.97	0.16	0.17	0.11	0.80	132.05	0.15	0.02	0.08	0.40	1.09
2	**89.65**	**0.27**	**-0.12**	0.12	***3.60***	**56.17**	**0.29**	0.15	**0.02**	***4.44***	***3.88***
3	293.35	0.19	0.03	0.13	0.16	171.29	0.17	**-0.02**	0.18	***2.42***	0.69
4	133.63	0.20	-0.03	0.06	0.89	101.34	0.20	0.24	0.30	1.16	1.22
5	573.30	0.19	-0.02	0.18	1.22	305.96	0.25	0.28	0.19	1.23	1.59
6	235.76	0.19	0.08	0.14	0.26	118.42	0.16	0.09	0.07	0.45	0.98
7	3622.78	0.17	0.10	0.10	0.13	1405.96	0.15	0.06	0.15	0.14	0.26
8	2459.44	0.15	0.04	0.08	0.32	1084.08	0.11	0.03	0.16	0.78	0.88
9	2310.05	0.17	0.11	0.15	0.62	1230.75	0.14	0.06	0.12	0.19	0.53
10	1040.77	0.14	0.05	0.11	0.55	541.43	0.15	0.13	0.23	0.47	2.38
11	275.02	0.16	0.04	0.13	0.40	166.93	0.14	0.12	0.13	0.46	2.37
12	428.91	0.18	0.02	0.20	1.50	253.76	0.15	0.01	0.06	0.55	1.30
13	244.40	0.22	0.13	0.10	1.14	139.65	0.21	0.27	0.33	1.81	1.27
14	904.78	0.22	0.09	0.07	0.99	466.45	0.19	0.10	0.10	0.32	1.06
15	511.32	0.23	0.07	0.10	0.91	322.19	0.23	0.25	0.23	0.89	1.29
16	984.44	**0.11**	**0.01**	**-0.01**	***3.19***	454.63	0.13	0.07	0.17	0.38	2.25
17	273.02	0.17	0.13	0.13	0.51	147.53	0.17	0.26	0.18	1.65	2.25
18	2427.84	0.20	0.06	0.11	0.12	871.31	0.15	0.10	0.28	1.23	1.11
19	4195.56	0.16	-0.09	0.11	1.62	1555.02	0.16	0.02	0.10	0.48	0.57
20	833.33	0.16	-0.04	0.05	1.12	368.48	0.15	-0.01	0.11	0.64	0.54
21	2237.82	0.13	0.19	0.10	1.53	929.32	0.11	0.02	0.08	0.84	***3.07***
22	5014.35	0.18	0.06	0.06	0.28	1943.34	0.18	0.20	0.16	0.39	1.02
23	1176.24	0.21	0.14	0.08	1.06	440.95	0.19	0.11	0.00	1.33	1.32
24	2398.37	0.17	-0.03	0.13	0.62	1219.86	0.18	0.22	0.25	0.67	0.92
25	2222.21	0.17	0.08	0.14	0.31	1339.62	0.19	0.23	0.09	1.35	1.11
26	5534.18	0.16	0.10	0.02	1.16	2222.26	0.15	0.10	0.11	0.32	0.54

Numbers written in bold and italic character indicate the discordant sites; those written in bold character represent the particular values of *L*-moments that possibly caused the discordancy.

Table 3. Weighted *L*-moments and corresponding model parameters

			The weighted L-moments after removing the discordant sites				Kappa parameters				Copula parameter
		Removed discordant sites	$\bar{l}_1$	$\bar{t}_2$	$\bar{t}_3$	$\bar{t}_4$	h	κ	α	β	m
Univariate	Volume	2 and 16	1.00	0.18	0.06	0.11	0.1053	0.2258	0.3196	0.8569	-
	Peak	2	1.00	0.17	0.12	0.15	-0.2887	0.0024	0.2150	0.9090	-
	Peak	2 and 3	1.00	0.17	0.12	0.15	-0.2370	0.0061	0.2197	0.9016	-
Bivariate	Volume	2 and 21	1.00	0.18	0.05	0.10	0.1200	0.2411	0.3250	0.8552	1.8280
	Peak	2 and 21	1.00	0.17	0.12	0.15	-0.3089	-0.0061	0.2138	0.9098	

Table 4. Homogeneity results

	Discordant sites	$V_{\|\|}$obs	μ_{sim}	σ_{sim}	$H_{\|\|}$	Decision
Bivariate (V, Q)	2 and 21	0.0296	0.0251	0.0029	1.5962	possibly homogeneous
Univariate V	2 and 16	0.0231	0.0218	0.0031	0.4279	homogeneous
Univariate Q	2	0.0321	0.0232	0.0035	2.5664	heterogeneous
	2 and 3	0.0329	0.0228	0.0035	2.8681	heterogeneous

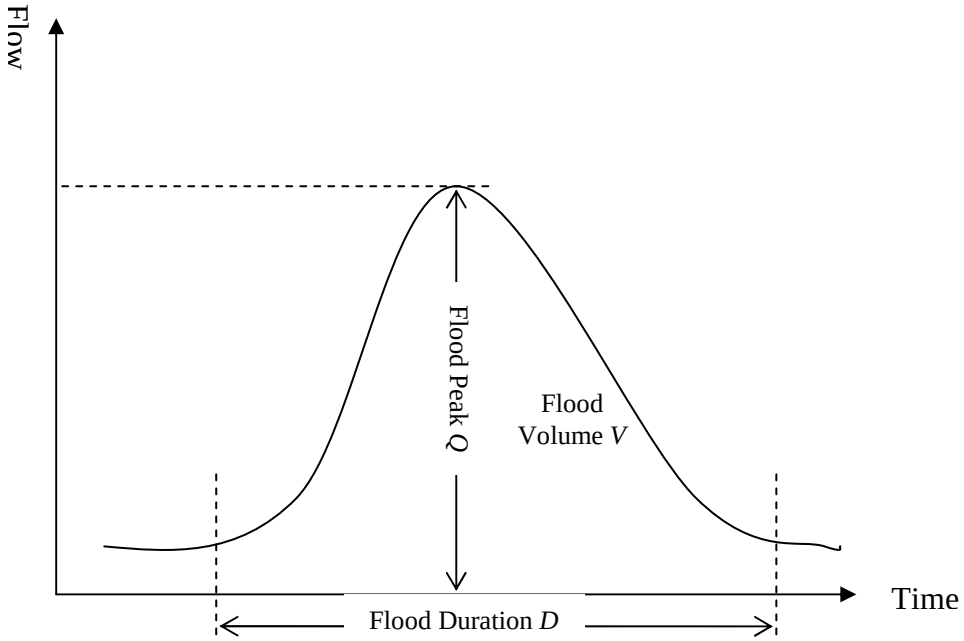


Figure 1. Typical flood hydrograph

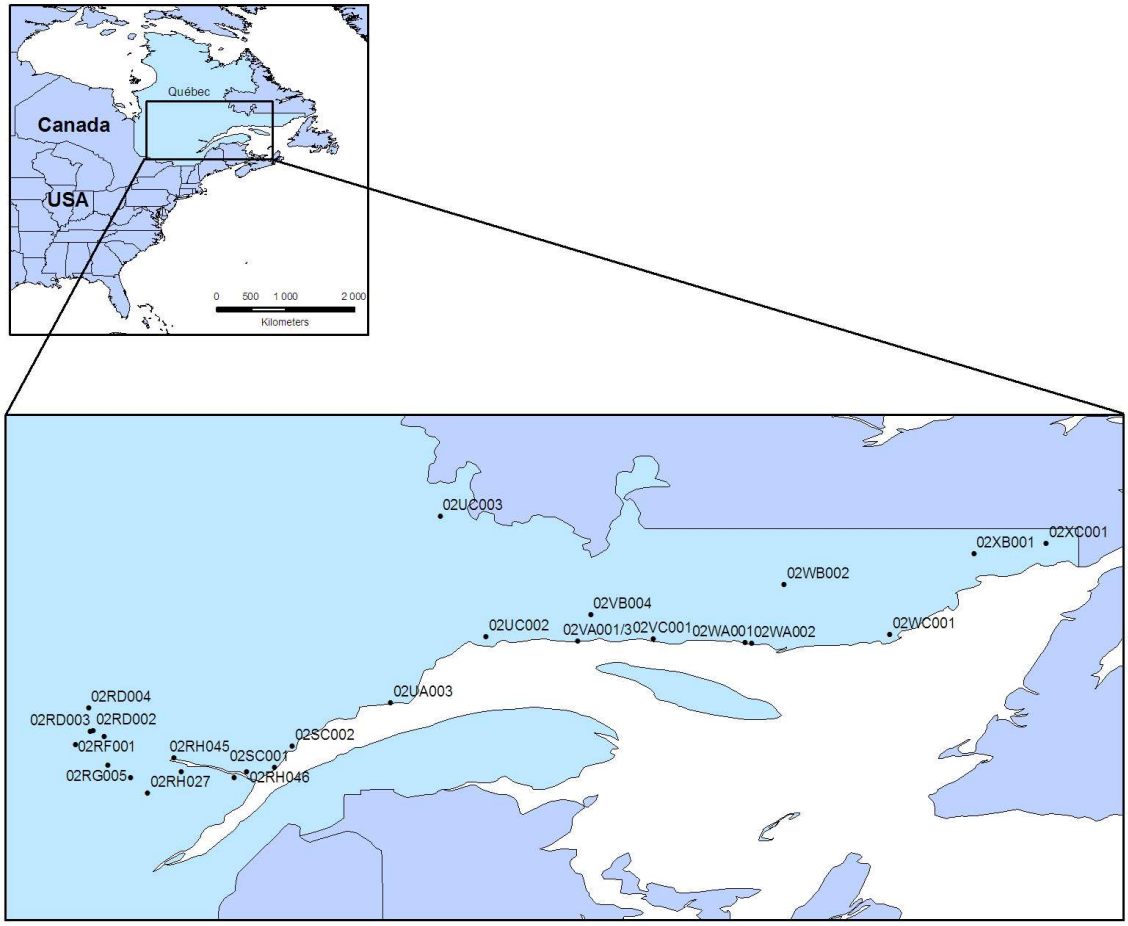


Figure 2. Geographical chart of the location of the sites

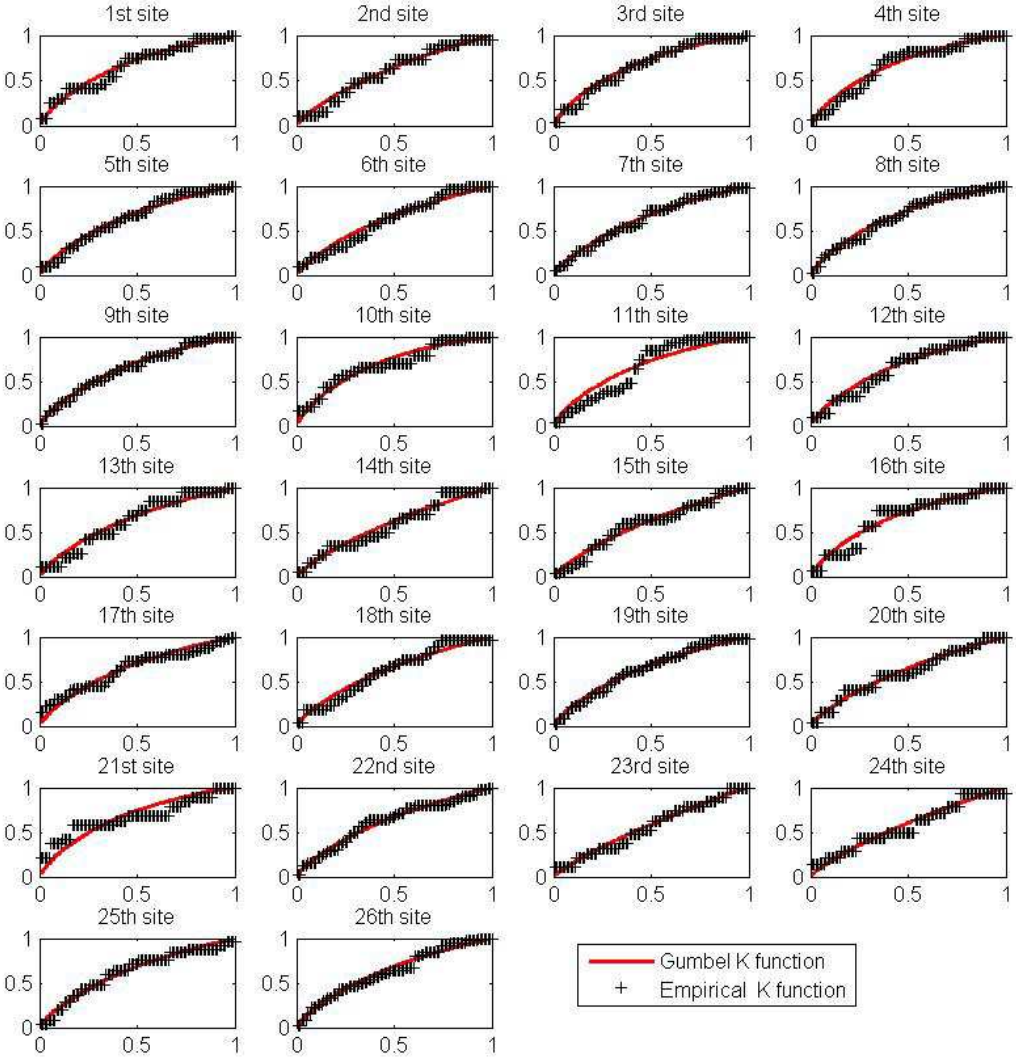
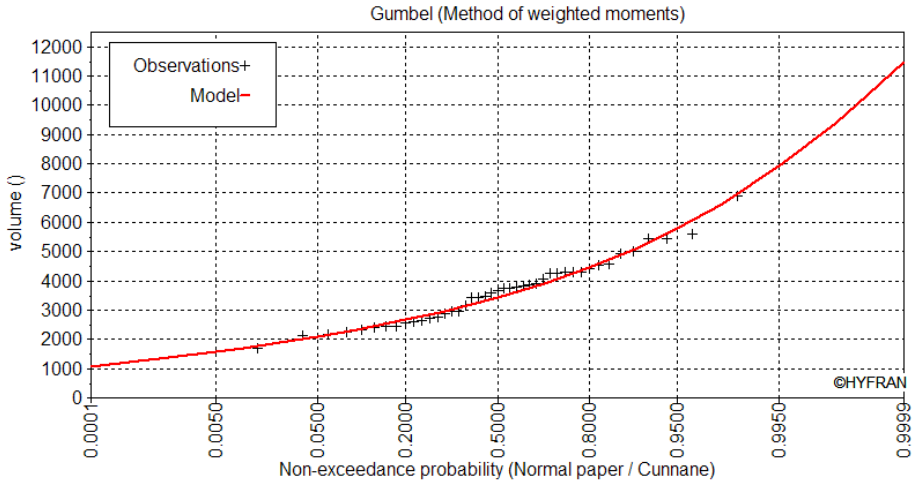
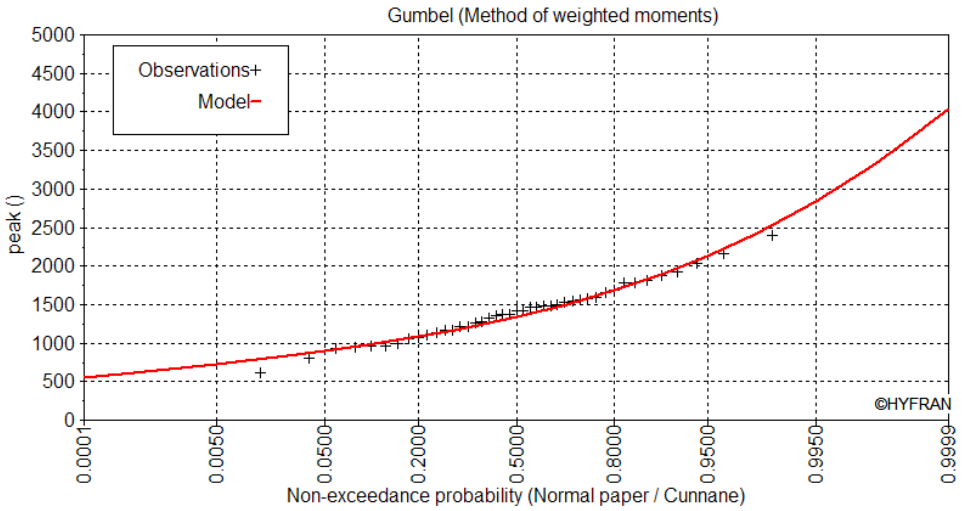


Figure 3. Fitting Gumbel logistic copula to each site of the data

(a)



(b)



(c)

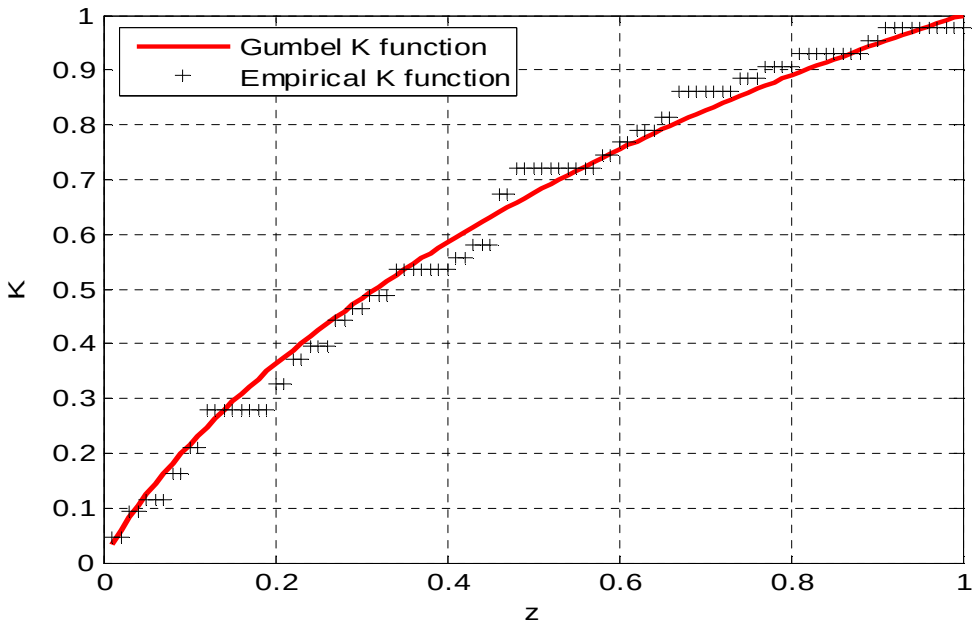


Figure 4. Distribution of volume (a), peak (b) and joint volume and peak (c) for the data of station 02RF001 (the 7th station)