Multivariate hydrological frequency analysis using copulas

Anne-Catherine Favre,[1] Salaheddine El Adlouni,[1] Luc Perreault,[2] Nathalie Thiémonge,[3] and Bernard Bobée[1]

Received 4 July 2003; revised 3 October 2003; accepted 21 October 2003; published 8 January 2004.

[1] This article presents the modeling of multivariate extreme values using copulas. Our approach allows us to model the dependence structure independently of the marginal distributions, which is not possible with standard classical methods. The methodology has been applied on two different problems in hydrology. The first application is concerned with the combined risk in the framework of frequency analysis. Four copulas have been tested on peak flows from the watershed of Peribonka in Québec, Canada. The second application relates to the joint modeling of peak flows and volumes. Three copulas have been applied to the watershed of the Rimouski River in Québec, Canada. This approach using copulas is promising since it allows us to take into account a wide range of correlation which can happen in hydrology. *INDEX TERMS:* 1821 Hydrology: Floods; 1860 Hydrology: Runoff and streamflow; 1869 Hydrology: Stochastic processes; *KEYWORDS:* frequency analysis, extreme values, bivariate distribution, copulas

Citation: Favre, A.-C., S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée (2004), Multivariate hydrological frequency analysis using copulas, *Water Resour. Res.*, *40*, W01101, doi:10.1029/2003WR002456.

1. Introduction

[2] In many applied statistical fields, such as hydrology, the analysis of multivariate events is of particular interest. For instance, design of hydropower dam requires the evaluation of the risk associated with peak discharges at the future dam site. In many cases, these peak discharges are the result of a combination of the rivers tributaries discharges upstream the location of interest. In such a case, the dependencies between all the quantities, which define the peak discharges should be taken into account. This necessarily involves a multivariate approach. Using a simple univariate approach could lead to severe underestimation of the risk associated to a given event [*Raynal-Villasenor and Salas*, 1987; *Bruneau et al.*, 1994]. Complex hydrological events such as floods and storms always appear to be multivariate events that are characterized by a few correlated random variables (peak, volume, duration, etc.). Therefore single-variable hydrological frequency analysis can only provide limited assessment of these events [*Yue et al.*, 2001].

[3] A large range of techniques has been developed and applied in hydrology to perform univariate analyses of extreme events [see, e.g., *Stedinger et al.*, 1993]. However, multivariate analysis of such random variables is rarely performed, in part because the very limited number of multivariate models available are not well suited to represent extreme values. The normal model has long dominated the statistical study of multivariate distributions. For example, leading studies on multivariate analysis, such as those of *Anderson* [1958] and *Johnson and Wichern* [1988], focus exclusively on the multivariate normal and related distributions that can be derived from normal distributions, including multivariate extensions of Student's t and Fischer's F distributions. Multivariate normal distributions are appealing because both the conditional and the marginal distributions are also normal. More recent texts on multivariate analysis, such as that by *Krzanowski* [1988], have begun to recognize the need for examining alternatives to the normal distribution setup. An extensive literature in statistics deals with nonnormal multivariate distributions [see, e.g., *Johnson and Kotz*, 1972; *Johnson et al.*, 1997]. However, many multivariate distributions have been developed as immediate extensions of univariate distributions, examples being the bivariate Pareto, bivariate gamma, etc. The drawbacks of these types of distributions are that (1) the same family is needed for each marginal distribution, (2) extensions to more than just the bivariate case are not clear, and (3) parameters of the marginal distributions are also used to model the dependence between the random variables. In hydrology the most used multivariate distributions are the multivariate normal, bivariate exponential [*Favre et al.*, 2002], bivariate gamma [*Yue et al.*, 2001], and bivariate extreme value distributions [*Adamson et al.*, 1999]. In the case of the multivariate normal the measure of dependence is summarized in the correlation matrix. In most cases, the use of a multivariate normal distribution is not appropriate to model maximum discharges because marginal distributions are asymmetric and have a heavy tail. Also, the dependence structure is generally different from the Gaussian case described by Pearson's correlation coefficient. Furthermore, in the case of more complex marginal distributions, such as finite mixtures of distributions, which are now widely used in practical modeling to represent heterogeneous phenomena [*Titterington et al.*, 1985; *West*, 1992; *Robert*, 1996] it is not possible to use standard multivariate distributions. In the first case study treated herein, we faced this type of problem since one of the samples involved in the analysis contains observations generated from two distinct processes.

[1] Chaire en hydrologie statistique, INRS, Eau-Terre et Environnement, Université du Québec, Sainte-Foy, Quebec, Canada.
[2] Institut de Recherche d'Hydro-Québec, Varennes, Quebec, Canada.
[3] Conception des aménagements de production, Hydro-Québec, Montréal, Quebec, Canada.

0043-1397/04/2003WR002456$09.00

[4] A construction of multivariate distributions that does not suffer from the drawbacks mentioned above is based on the notion of copulas [*Sklar*, 1959]. A copula is very useful to implement efficient algorithms for simulating joint distributions in a more realistic way. In fact, copulas are able to model the dependence structure independently of the margin distributions. It is then possible to build multidimensional distributions with different margins, the structure of dependence being mathematically formalized through the copula. The crucial step in the modeling process is the choice and the adjustment of the copula function which best fits the data. Copulas have been widely used in the financial domain in order to determine the value at risk [see, e.g., *Embrechts et al.*, 2002, 2003; *Bouyé et al.*, 2000]. Other fields of applications involve lifetime data analysis [*Bagdonavicius et al.*, 1999] and actuarial science [*Frees and Valdez*, 1998]. However the use of copulas in the hydrologic domain is still a marginal phenomenon. A few authors use a particular bivariate distribution, the Farlie-Gumbel-Morgenstern distribution, but without referring to copulas [*Singh and Singh*, 1991; *Long and Krzysztofowicz*, 1992]. *De Michele and Salvadori* [2003] model different combinations of rainfall depth and duration using copulas. In their case the random variables show negative dependence. *Wang* [2001] proposes a Bayesian estimation for the parameters of copula of extreme values but without doing a complete application.

[5] In this paper we present two applications in hydrology using copulas. The first application concerns a study about the combined risk in Peribonka. In this case, copulas are the only realistic way to handle the problem since the involved marginals are different and non standard. The second application deals with the bivariate analysis of the volume and flow in Rimouski. Here again the involved marginals are different and a classical approach can not be used.

[6] The next section is devoted to the general theory about copulas as these types of multivariate distributions are not familiar in hydrological literature. We present the definition of copulas and their main properties (section 2.1), the main types of copulas (section 2.2), the parameters estimation (section 2.3), and the simulation of copulas (section 2.4). Section 3 deals with the first application, namely, the problem definition (section 3.1), the modeling (section 3.2), and the obtained results (section 3.3). Section 4 shows the second application. Section 5 is devoted to the conclusions and prospects for further research.

2. General Theory About Copulas

[7] The theory about copulas can be found in general textbooks such as those of *Nelsen* [1999] and *Joe* [1997]. This section offers an overview of the main concepts.

2.1. Properties of Copulas

[8] To define a copula, consider p uniform $\mathcal{U}(0,1)$ random variables $U_1,\cdots,U_p$. Unlike many applications we do not assume that $U_1,\cdots,U_p$ are independent; to the contrary they are assumed to be related. The relationship between these random variables is described through their joint distribution function as

$$C(u_1,\cdots,u_p) = \Pr(U_1 \le u_1,\cdots,U_p \le u_p).$$

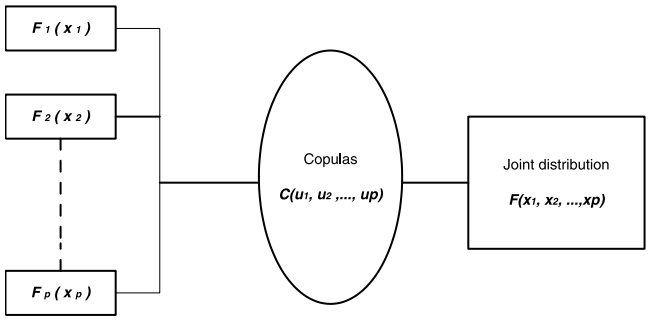

Figure 1. Representation of a copula.

Here, we call the function C a copula. To complete the construction, we select arbitrary marginal distribution functions $F_1(x_1),\cdots, F_p(x_p)$. Then the function

$$C(F_1(x_1),\cdots,F_p(x_p)) = F(x_1,\cdots,x_p) \tag{1}$$

defines a multivariate distribution function, evaluated at $x_1,\cdots, x_p$. Figure 1 represents schematically the notion of copula.

[9] In the copula model defined in equation (1) we can integrate different families of probability distributions for each outcome. This is the main advantage of this approach compared to standard multivariate models used in practice. It is easy to check from the construction in equation (1) that F is a multivariate distribution. *Sklar* [1959] established the converse. He showed that any multivariate distribution function F can be written in the form of equation (1), that is using a copula representation. Sklar also showed that if the marginal distributions are continuous, then there is a unique copula representation. From Sklar's theorem, we see that for continuous multivariate distribution functions the univariate margins and the multivariate dependence structure can be separated, and the dependence structure can be represented by a copula. In the remaining of the article we limit the discussion to the bivariate case for simplicity reasons.

[10] *Schweizer and Wolff* [1981] established that the copula accounts for all the dependence between two random variables, X_1 and X_2, in the following sense. Consider g_1 and g_2, two strictly increasing functions (but otherwise arbitrary) over the range of X_1 and X_2. Then the transformed variables $g_1(X_1)$ and $g_2(X_2)$ have the same copula as X_1 and X_2. Thus, as stated by *Frees and Valdez* [1998], the manner in which X_1 and X_2 "move together" is captured by the copula, regardless of the scale in which each variable is measured.

[11] Schweizer and Wolff also showed that two standard nonparametric correlation measures could be expressed solely in terms of the copula function. These are Spearman's correlation coefficient, defined by

$$\begin{aligned}\rho_s &= \rho(F_1(x_1), F_2(x_2)) \\ &= 12\int\int F_1(x_1)F_2(x_2)dF(x_1,x_2) - 3 \\ &= 12\int\int uvdC(u,v) - 3\end{aligned} \tag{2}$$

Table 1. Bivariate Archimedean Copulas and Their Generators

Family	Generator $\varphi(t)$	Parameter α	Bivariate Copula $C_\varphi(u_1, u_2)$
Independence	$-\ln t$	–	$u_1 u_2$
Clayton [1978], *Cook and Johnson* [1981], and *Oakes* [1982]	$t^{-\alpha} - 1$	$\alpha > 0$	$(u_1^{-\alpha} + u_2^{-\alpha} - 1)^{-1/\alpha}$
Gumbel [1960] and *Hougaard* [1986]	$(-\ln t)^\alpha$	$\alpha \geq 1$	$\exp\{-[(-\ln u_1)^\alpha + (-\ln u_2)^\alpha]^{1/\alpha}\}$
Frank [1979], *Nelsen* [1986], and *Genest* [1987]	$\ln\left(\frac{e^{\alpha t}-1}{e^{\alpha}-1}\right)$	$\alpha \neq 0$	$\frac{1}{\alpha}\ln\left(1 + \frac{(e^{\alpha u_1}-1)(e^{\alpha u_2}-1)}{e^{\alpha}-1}\right)$

and Kendall's correlation coefficient, defined by

$$\begin{aligned}\tau &= \Pr((X_1 - X_1^*)(X_2 - X_2^*) > 0) \\ &\quad - Pr((X_1 - X_1^*)(X_2 - X_2^*) < 0) \\ &= 2Pr((X_1 - X_1^*)(X_2 - X_2^*) > 0) - 1 \\ &= 4\int FdF - 1 \\ &= 4\int CdC - 1. \end{aligned} \tag{3}$$

For these expressions, we assume that X_1 and X_2 have a jointly continuous distribution function. Further, the definition of Kendall's τ uses an independent copy of (X_1, X_2) namely (X_1^*, X_2^*) to define the measure of concordance. These two important properties can be used to estimate parameters of several copulas. The widely used Pearson correlation coefficient, $\mathrm{Cov}(X_1, X_2)/(\mathrm{Var}(X_1)\ \mathrm{Var}(X_2))^{1/2}$ depends not only on the copula but also on the marginal distributions. Thus this measure is affected by nonlinear changes of scale. Moreover it is not invariant to monotonic transformations contrary to Spearman's and Kendall's measures.

[12] Note that *Blest* [2000] recently proposed a new nonparametric measure of the dependence between two random variables. This coefficient gives more weight to observed difference in the first ranks. A detailed study of this coefficient is given by *Genest and Plante* [2003].

[13] It is worth to notice that each copula is bounded by the so-called Frchet-Hoeffding bounds, so that

$$\begin{aligned}&\max(F_1(x_1) + \cdots + F_p(x_p) - 1, 0) \leq F(x_1, \cdots, x_p) \\ &\quad \leq \min(F_1(x_1), \cdots, F_p(x_p)).\end{aligned}$$

These bounds correspond to the situations where the two random variables are almost surely monotone increasing and decreasing functions of one another. The proof of this theorem is given by *Nelsen* [1999, theorem 2.2.3 p. 8].

2.2. Types of Copulas

2.2.1. Elliptical Copulas

[14] Copulas related to elliptical distributions are very useful in practical applications since they have several properties of the multivariate normal distribution. The most well known elliptical copulas are the multivariate gaussian copula and the multivariate Student copula.

Definition 2.1. Let ρ be a symmetric, positive definite matrix with $\mathrm{diag}\rho = 1$ and ϕ_ρ the standardized multivariate normal distribution with correlation matrix ρ. The multivariate gaussian copula is then defined as follows

$$C(u_1, \cdots, u_p; \rho) = \phi_\rho\left(\phi^{-1}(u_1), \cdots, \phi^{-1}(u_p)\right)$$

Definition 2.2. Let ρ be a symmetric, positive definite matrix with $\mathrm{diag}\rho = 1$ and $T_{\rho,\nu}$ the standardized multivariate Student's distribution with ν degrees of freedom and correlation matrix ρ. The multivariate Student's copula is then defined as follows:

$$C(u_1, \cdots, u_p; \rho, \nu) = T_{\rho,\nu}\left(t_\nu^{-1}(u_1), \cdots, t_\nu^{-1}(u_p)\right)$$

where t_ν^{-1} is the inverse of the univariate Student's distribution.

[15] Note that a generalization of Definitions 2.1 and 2.2 is possible introducing an asymmetry by indexing the copula by a matrix of dependence parameters [*Joe*, 1997]. There are few works that focus on elliptical copulas. However, they could be very attractive. *Jorgensen* [1997] and *Song* [2000] proposed a multivariate extension of dispersion models with the Gaussian copulas.

2.2.2. Archimedean Copulas

[16] Archimedean copulas originally appeared not in statistics, but rather in the study of probabilistic metric spaces, where they were studied as part of the development of a probabilistic version of the triangle inequality [see *Schweizer*, 1991]. *Genest and MacKay* [1986a] define Archimedean copulas as the following:

$$C(u_1, \cdots, u_p) = \begin{cases} \varphi^{-1}\left(\varphi(u_1) + \cdots + \varphi(u_p)\right) & \text{if } \sum_{j=1}^{p}\varphi(u_j) \leq \varphi(0) \\ 0 & \text{otherwise} \end{cases}.$$

where $\varphi(u)$ a $\mathcal{C}^2$ function with $\varphi(1) = 0$, $\varphi'(u) < 0$ (φ is decreasing), $\varphi''(u) > 0$ (φ is convex) for all $0 \leq u \leq 1$ and the first p derivatives of φ are of alternating signs. $\varphi(u)$ is called the generator of the copula. Archimedean copulas play an important role because they present several desired properties (C is symmetric, associative,...). Moreover, in Archimedean copulas the computation of measures of dependence is simplified. For example, equation (3) for Kendall's tau reduces to

$$\tau = 1 + 4\int_0^1 \frac{\varphi(u)}{\varphi'(u)}du.$$

Note that a limitation of Archimedean copulas is that they are symmetric in their arguments. Extensions are possible in which this symmetry condition is evacuated [see *Joe*, 1997].

[17] Table 1 shows that different choices of generator yield several important bivariate families of copulas. A generator uniquely determines (up to a scalar multiple) an Archimedean copula.

[18] Note that throughout the article, the copulas number 2 and 4 will be related to Clayton and Frank copulas respectively. We can also cite the copulas of *Ali et al.*

[1978], *Cuadras and Augé* [1981], *Galambos* [1975], *Hüsler and Reiss* [1989], and *Genest and Ghoudi* [1994], which also belong to the class of Archimedean copulas.

2.2.3. Copulas With Quadratic Section

[19] The Farlie-Gumbel-Morgenstern family of copulas belong to the class of copulas with quadratic section. In the bivariate case this copula is defined as

$$C(u_1, u_2) = u_1 u_2 + \alpha u_1 u_2 (1 - u_1)(1 - u_2)$$

with $\alpha \in [-1,1]$. The family was discussed by *Morgenstern* [1956], *Gumbel* [1958], and *Farlie* [1960]. However it seems that the earliest publication is that of *Eyraud* [1938]. Because of their simple analytical form, Farlie-Gumbel-Morgenstern distributions have been widely used in modeling. Note that Farlie-Gumbel-Morgenstern copula does not belong to the family of Archimedean copulas since they are not associative.

2.3. Parameter Estimation

[20] Let θ be the $K \times 1$ vector of parameters (comprising the dependence parameters and also the marginal distributions parameters) to be estimated and Θ the parameter space. The log likelihood for observation i is denoted $l_i(\theta)$. Given n independent observations, we get

$$l(\theta) = \sum_{i=1}^{n} l_i(\theta).$$

Applied to equation (1), the expression of the log likelihood becomes in the case of copulas

$$l(\theta) = \sum_{i=1}^{n} \ln c(F_1(x_1^i), \cdots, F_p(x_p^i)) + \sum_{i=1}^{n} \sum_{j=1}^{p} \ln f_j(x_j^i).$$

$\hat{\theta}_{\mathrm{ML}}$ the maximum likelihood estimator satisfies

$$l(\hat{\theta}_{ML}) \geq l(\theta) \quad \forall \theta \in \Theta.$$

The previous method, which is called the exact maximum likelihood method (EML) could be computational intensive in the case of high dimensional distribution, because it requires to jointly estimate the parameters of the margins and the parameters of the dependence structure. However, the copula representation splits the parameters into specific parameters for marginal distributions and common parameters for the dependence structure. The log likelihood could then be written as

$$l(\theta) = \sum_{i=1}^{n} \ln c(F_1(x_1^i; \theta_1), \cdots, F_p(x_p^i; \theta_p); \alpha) + \sum_{i=1}^{n} \sum_{j=1}^{p} \ln f_j(x_j^i; \theta_j)$$

with $\theta = (\theta_1, \cdots, \theta_p, \alpha)$. We can also perform the estimation of the univariate marginal distributions in a first step

$$\hat{\theta}_j = \arg\max \sum_{i=1}^{n} \ln f_j(x_j^i; \theta_j)$$

and then estimate α given the previous estimates

$$\hat{\alpha} = \arg\max \sum_{i=1}^{n} \ln c(F_1(x_1^i; \hat{\theta}_1), \cdots, F_p(x_p^i; \hat{\theta}_p); \alpha).$$

This two-step method is called the method of inference functions for margins (IFM) method.

[21] Note that following the definition of empirical copulas introduced by *Deheuvels* [1979], *Genest and Rivest* [1993] have developed a nonparametric method to identify the copula in the Archimedean case. This method has been further improved by *Barbe et al.* [1996].

2.4. Simulation of Copulas

[22] *Genest and MacKay* [1986b] proposed a general algorithm to simulate a copula of the Archimedean family. They introduced the idea of simulating the full distribution of $(X_1, \cdots, X_p)$ by recursively simulating the conditional distribution of X_j given X_1. The algorithm is summarized by *Lee* [1993] as the following.

[23] 1. Generate $U_1, \cdots, U_p$ independent $\mathcal{U}(0,1)$ random numbers.

[24] 2. Set $X_1 = F_1^{-1}(U_1)$ and $c_0 = 0$.

[25] 3. For $j = 2, \cdots, p$ recursively calculate X_j as the solution of

$$U_j = F_j(X_j | x_1, \cdots, x_{j-1}) = \frac{\varphi^{-1(j-1)}\{c_{j-1} + \varphi[F_j(x_j)]\}}{\varphi^{-1(j-1)}(c_{j-1})}$$

where $c_j = \varphi[F_1(x_1)] + \cdots + \varphi[F_j(x_j)]$.

3. Application: Flow Combination

3.1. Problem Definition

[26] The watershed of the Peribonka river is located in the Canadian province of Quebec in the hydrographical region 06 between latitudes 48° 45′ and 52° North and longitudes 70° and 72°W. The hydroelectrical works actually under study by Hydro-Quebec are located on the Peribonka river, 151.8 km upstream from the outflow in the St-Jean Lake. Hydro-Québec is a public company that produces, transmits and distributes electricity throughout the province of Québec. A rock fill dam is built about 200 m upstream from the confluence of the Manouane river. Figure 2 shows the localization of the watershed.

[27] The planned hydroelectrical works (site PER-3D) consists of a run-of-river power station. The watershed area at the power station is 19 450 km^2 while the intermediate watershed has an area of 3 133 km^2. The design of the hydroelectrical works is planned in order not to influence the management of the existing dams of Alcan Inc., an international company providing aluminum, on the Peribonka river.

[28] The peak flows at Peribonka (site PER-3D) are the combination of the outflows from the upstream site Chute-des-Passes and the peak flows from the intermediate watershed (area of 3 133 km^2). The resulting peak flow can be written as $Z = X + Y$, where X is the random variable representing the flow at Chute-des-Passes and Y at the intermediate watershed. Figure 3 illustrates the problem.

[29] The aim is to estimate the annual peak flow of a given return period taking into account the dependence between the flows in order to design the hydroelectrical works (PER-3D).

3.2. Modeling

[30] The available data are the annual maximum peak flow at Chute-des-Passes from 1960 to 2001 and the annual

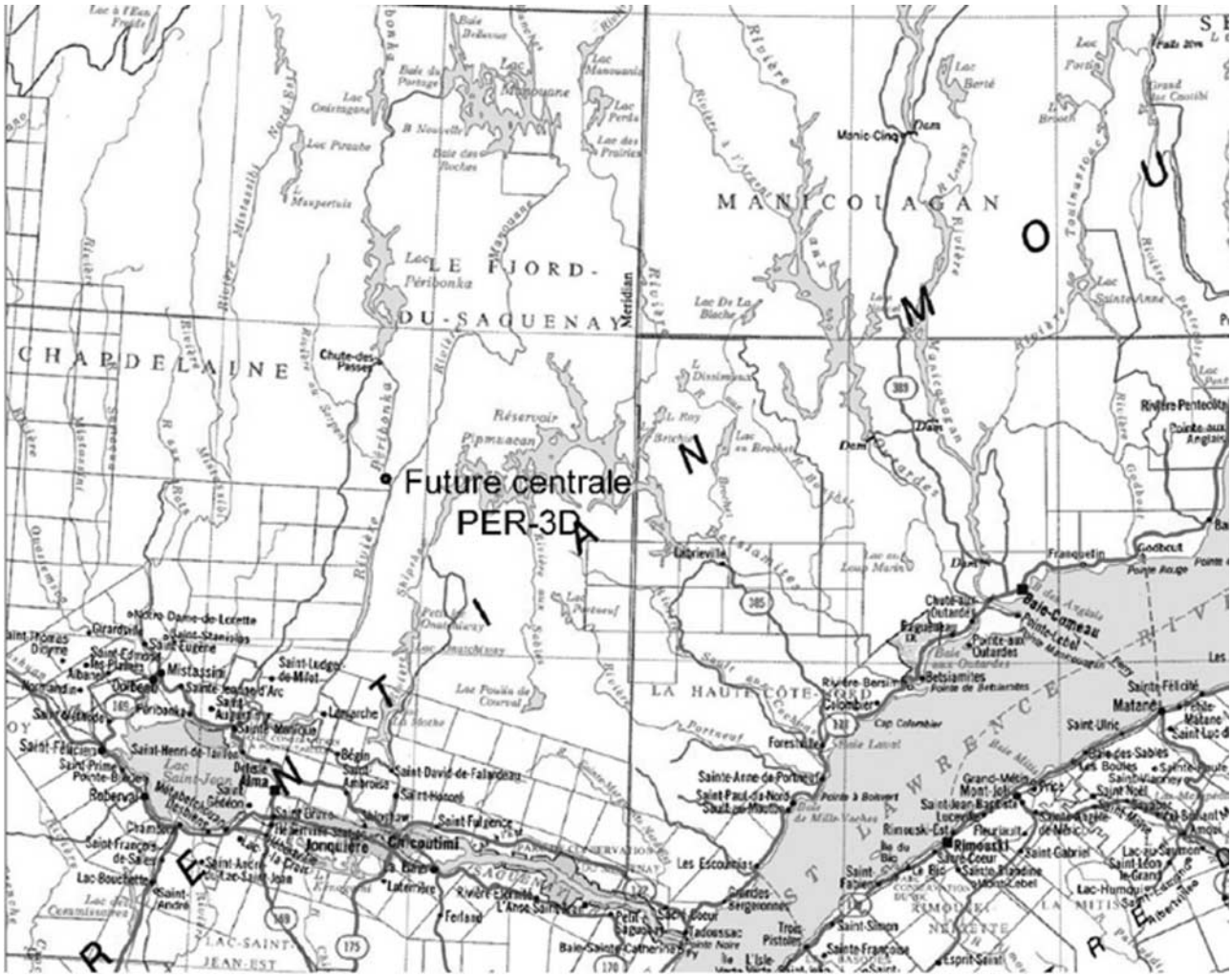

Figure 2. Watershed map.

maximum peak flow at the intermediate watershed from 1979 to 2002. Figure 4 illustrates the time series outflow from Chute-des-Passes.

[31] Two flow regimes can be distinguished: high flows corresponding to wet years and moderate flows corresponding to dry years. High flows result from spillovers. Figure 5 shows the empirical cumulative distribution at Chute-des-Passes which clearly exhibits a heterogeneity.

[32] The spillovers can not be explained only by meteorological reasons but specialists from Hydro-Québec think that such kind of spillovers can happen again in the future. Therefore we must take into account such additional uncertainty in the analysis. To do so we chose to model this data with a mixture of two univariate normal proability distribution. Such model is especially designed to represent hetero-

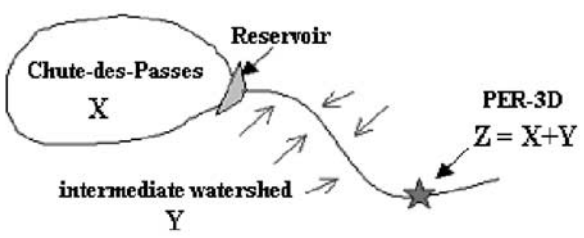

Figure 3. Problem of combined risk at Peribonka.

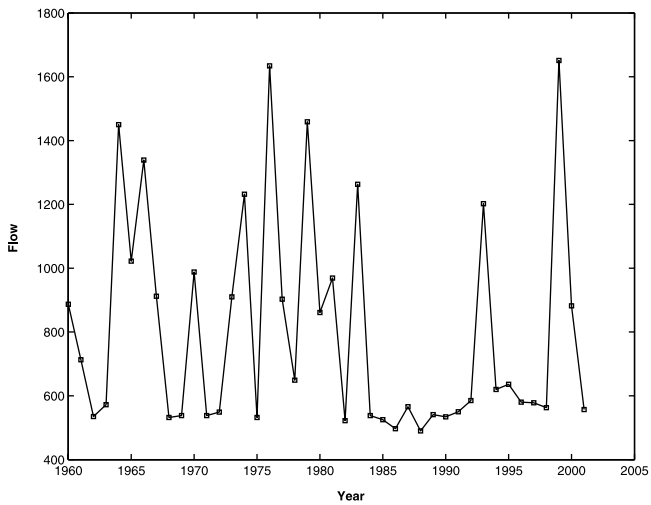

Figure 4. Time series of annual peak flow at Chute-des-Passes.

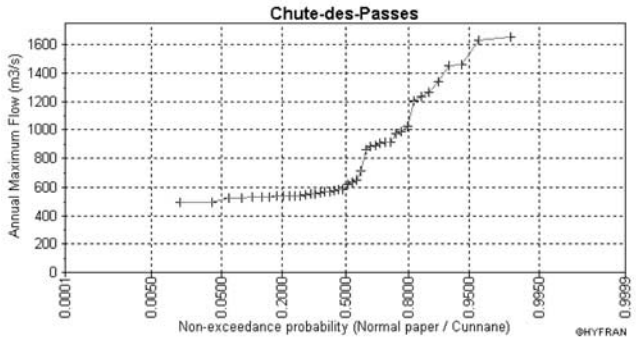

Figure 5. Empirical cumulated distribution at Chute-des-Passes.

geneous observations [*Robert*, 1996]. The probability density function of such a mixture is expressed as follows:

$$X \sim \pi\mathcal{N}(\mu_1, \sigma_1) + (1 - \pi)\mathcal{N}(\mu_2, \sigma_2),$$

where $\mathcal{N}(\mu, \sigma)$ stands for the normal density function and π is the mixture proportion which represents the relative frequency of occurrence of process 1 (a normal distribution with mean μ_1 and standard deviation σ_1). The parameters have been estimated using a Bayesian approach (see *Perreault* [2003] for more details). Gibbs sampling was used to approximate the posterior density of each parameter. The expected values of these posterior distributions have been considered as estimates for the parameters. We obtained the following values:

$$\hat{\pi} = 0.48, \hat{\mu}_1 = 546\ \mathrm{m}^3/\mathrm{s}, \hat{\sigma}_1 = 25\ \mathrm{m}^3/\mathrm{s}, \hat{\mu}_2 = 1039\ \mathrm{m}^3/\mathrm{s},$$
$$\hat{\sigma}_2 = 314\ \mathrm{m}^3/\mathrm{s}.$$

Figure 6 shows the histogram with the superposed estimated density of the mixture model.

[33] The annual maximum peak flows at the intermediate watershed have been modelled using the Gamma distribution. Thus the density can be expressed as:

$$G(y; \beta, \lambda) = \frac{\beta \exp(-\beta y)(\beta y)^{\lambda - 1}}{\Gamma(\lambda)}$$

where $y \geq 0$ and $\beta, \lambda > 0$.

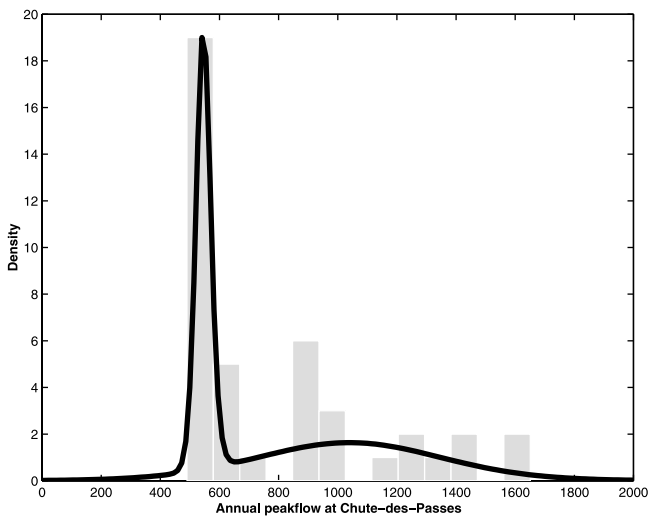

Figure 6. Histogram of the annual flow at Chute-des-Passes with the bivariate normal density.

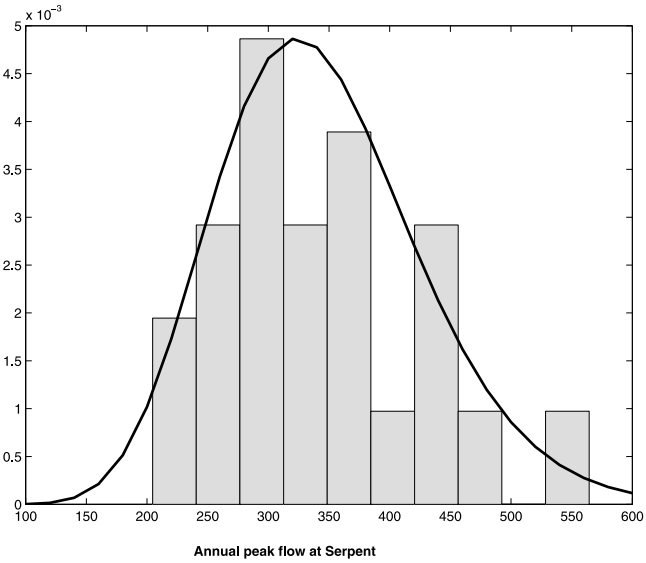

Figure 7. Histogram of the annual flow at Serpent with the gamma density.

[34] The parameters have been estimated using the method of maximum likelihood. The estimates obtained for each parameter are $\hat{\beta} = 16$, $\hat{\lambda} = 20$. Figure 7 shows the histogram and the obtained density. The plot in Figure 8 illustrates the nonexceedence probability with the corresponding asymptotic 95% confidence interval.

[35] The next step in our study is to define the bivariate distribution in order to take into account the dependence between the two flow series. Figure 9 shows a scatter plot of the pairs $(R_i/(n + 1), S_i/(n + 1))$, where R_i and S_i are respectively the rank of the flows at Serpent and Chute-des-Passes.

[36] This plot corresponds to the bivariate probability function associated with Deheuvels' empirical copula [*Deheuvels*, 1979]. We obtained the following values for the classical measures of dependence: $\rho = 0.2$, $\rho_S = 0.14$, $\tau = 0.16$. Standard bivariate distributions used in hydrology can not be applied in our case due to the complexity of one of the marginal distribution (in this case a mixture of distributions). The only possible method to deal with this problem in a formal way is to use copulas. We considered several types of copulas: the independence case, Farlie-Gumbel-

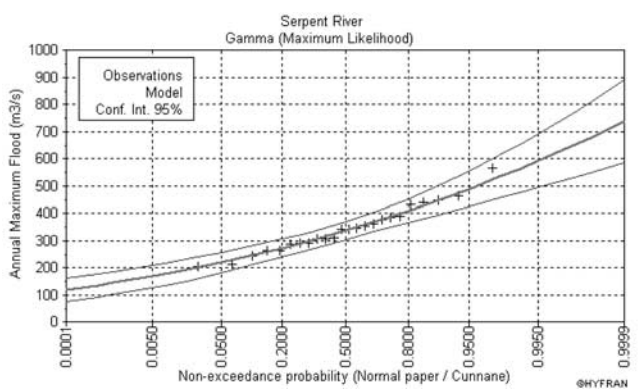

Figure 8. Observed and estimated flows with 95% confidence intervals for the intermediate watershed.

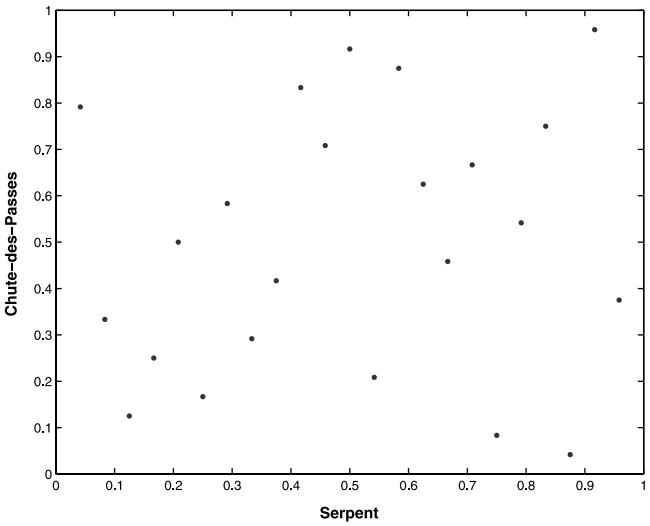

Figure 9. Bivariate probability function associated with Deheuvels' empirical copula.

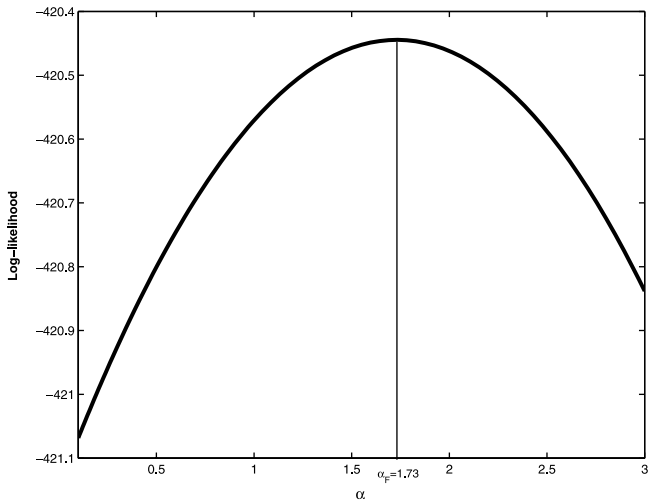

Figure 10. Log likelihood for Frank copula.

Morgenstern copula already used in hydrology [*Singh and Singh*, 1991], Frank and Clayton copulas. Their analytical expressions are given in Table 2.

[37] The last two copulas have been chosen for their simplicity (only one parameter to be estimated) and flexibility (they are able to model continuously a whole range of dependence between the lower Fréchet-Hoeffding bound copula, the independent copula and the upper Fréchet-Hoeffding bound copula). In our case, equation (1) can be rewritten as

$$F(x,y) = C(F_1(x), F_2(y)),$$

$$\text{with } F_1 \sim 0.48\mathcal{N}(546,25) + 0.52\mathcal{N}(1039,314)$$

$$F_2 \sim \mathcal{G}(16,20).$$

In the case of Farlie-Gumbel-Morgenstern and Clayton copulas, a very simple formula links the parameter of dependence (α) to Kendall's tau (τ) measure of dependence. We have that $\alpha_{FGM} = \frac{9}{2}\tau$ and $\alpha_C = \frac{2\tau}{1-\tau}$ [see, e.g., *Nelsen*, 1999]. We use these relations to estimate the corresponding parameters. For Frank copula, the method of inference functions for margins (IFM, see section 2.3) has been used. Figure 10 shows that the maximum of the log likelihood is well defined.

[38] We obtain the following values for α: $\alpha_F = 1.73$, $\alpha_C = 0.39$, $\alpha_{FGM} = 0.73$.

3.3. Results

[39] For all copulas defined in Table 2 a simulation of size 15 000 has been realized. Frank and Clayton copulas have been simulated using the general procedure described in section 2.5. For Farlie-Gumbel-Morgenstern's copula the method defined by *Johnson* [1987] has been adopted. Figure 11 shows the observations along with the simulated values.

Table 2. Type of Copulas Chosen for the Application

Name	$C(u_1,u_2)$
Independence	u_1u_2
Farlie-Gumbel-Morgenstern	$u_1u_2 + \alpha_{FGM}u_1u_2(1-u_1)(1-u_2)$
Frank	$\frac{1}{\alpha_F}\ln\left(1 + \frac{(\exp(\alpha_F u_1)-1)(\exp(\alpha_F u_2)-1)}{\exp(\alpha_F)-1}\right)$
Clayton	$(u_1^{-\alpha_C} + u_2^{-\alpha_C} - 1)^{-1/\alpha_C}$

[40] In each case the observations are situated in the scatter diagram of simulated values. A statistical test [*Genest and Rivest*, 1993] has been applied to choose between Frank and Clayton copulas. The coefficients of determination obtained are respectively of 97.2% and 96.3%, which shows that Frank copula seems to have a slightly better behavior. Figure 12 illustrates the joint distribution obtained in the case of Frank copula.

[41] With the simulated values, total discharges ($X + Y$) of a given return period have been determined using the empirical cumulative distribution. Figure 13 shows the plot of the obtained empirical cumulative distribution with the four models. Results are summarized in Table 3.

[42] In general no huge differences can be shown from a copula to another. The mean range is about 4%. Farlie-Gumbel-Morgenstern and Frank copulas show the highest similarity. For return periods between 20 and 1000 years, Frank copula gives the largest values. Even though the correlation is small, for more than a 2-years return period, the independence copula gives systematically smaller quantile values than the three other copulas. The highest range is obtained for 60 years with a magnitude of 5.5%. Therefore not taking into account the dependence may lead to under designing the hydroelectrical works, which increases the hydrological risk.

4. Application: Bivariate Frequency Analysis

4.1. Problem Definition

[43] Many hydrological engineering planning, design and management problems require a detailed knowledge of flood event characteristics, such as flood peak, volume and duration. Flood frequency analysis often focuses on flood peak values and hence provides a limited assessment of flood events. This second application concerns the bivariate frequency analysis of peak flow and volume of the Rimouski river. The watershed is situated in the south

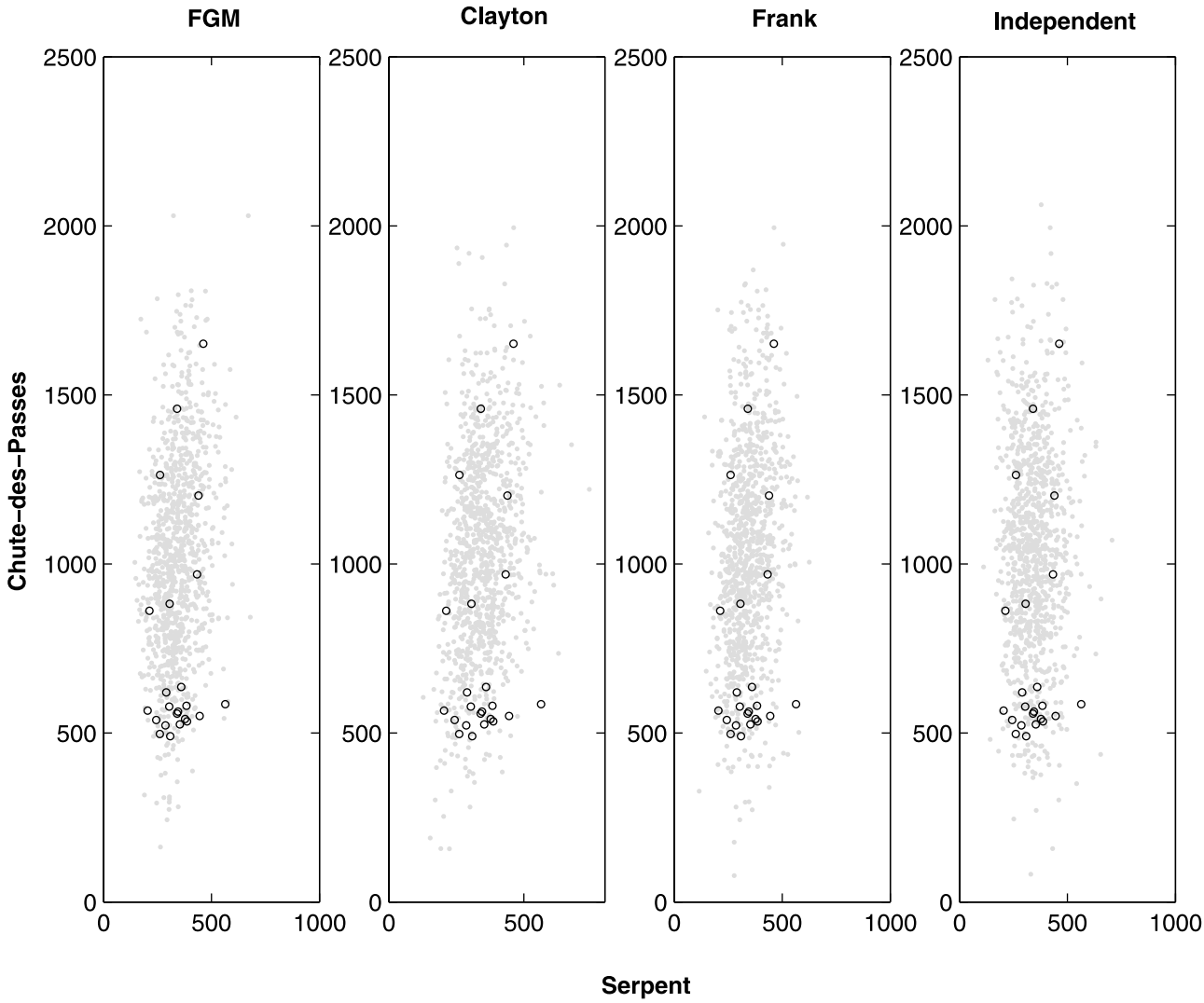

Figure 11. Observations versus simulations for the four types of copulas.

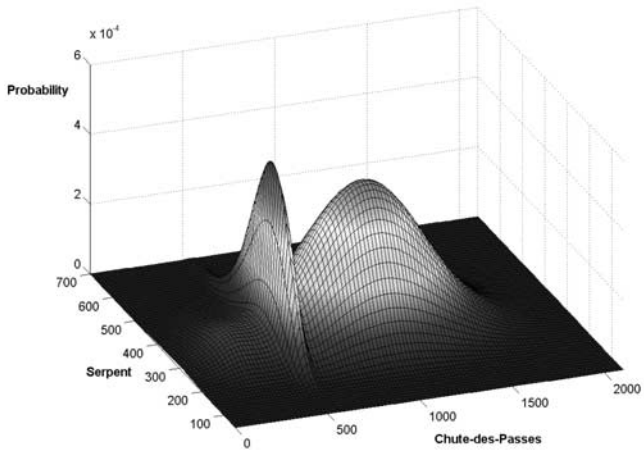

Figure 12. Joint distribution in the case of Frank copula.

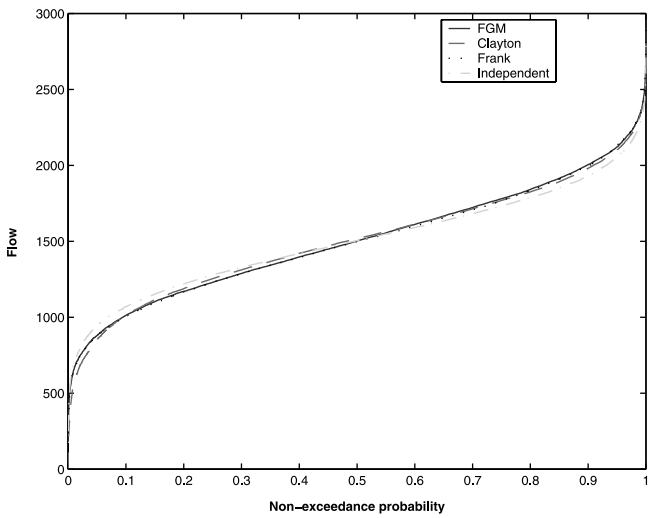

Figure 13. Empirical cumulative distribution in the cases of independence, Farlie-Gumbel-Morgenstern, Clayton, and Frank copulas.

Table 3. Flows of a Given Return Period for the Four Tested Models

Return Period T, years	Independence	FGM	Clayton	Frank
2	1497	1494	1510	1487
10	1918	1989	1974	1986
20	2038	2114	2101	2116
40	2114	2220	2208	2225
60	2165	2276	2260	2283
100	2243	2342	2330	2350
1000	2545	2630	2595	2635
10000	2706	2791	2756	2766

shore of the St. Laurent River. Its size is 1 610 km^2. The underlying data are the annual peak flows and volumes from 1963 to 1997. Figure 14 shows the flows and volumes as time series.

[44] Obviously, the two quantities are strongly correlated. Pearson's correlation coefficient is 0.76.

4.2. Modeling

[45] Several authors have considered the joint modeling of flows and volumes [see, e.g., *Singh and Singh*, 1991; *Adamson et al.*, 1999; *Yue et al.*, 2001]. In these studies, the considered marginal distributions for both random variables involved in the analysis were always identical (exponential, Gumbel, or gamma). However, the marginal distributions of flows and volumes often differ. This is the case in the application presented below.

[46] The annual maximum flows Q were best fitted by a Gumbel distribution (extreme value type I (EVI)). The parameters were estimated using the maximum likelihood method. We obtained $Q \sim EVI(223, 70)$. Figure 15 shows the results on a nonexceedance probability plot with the related 95% asymptotic confidence interval.

[47] For the annual maximum volumes V a gamma distribution was considered. This distribution is frequently used in hydrology for volumes [*Yue et al.*, 2001]. The parameters were estimated with the maximum likelihood method. We obtained $V \sim \mathcal{G}(16,32)$. Figure 16 shows the results on a nonexceedance probability plot with the related 95% asymptotic confidence interval.

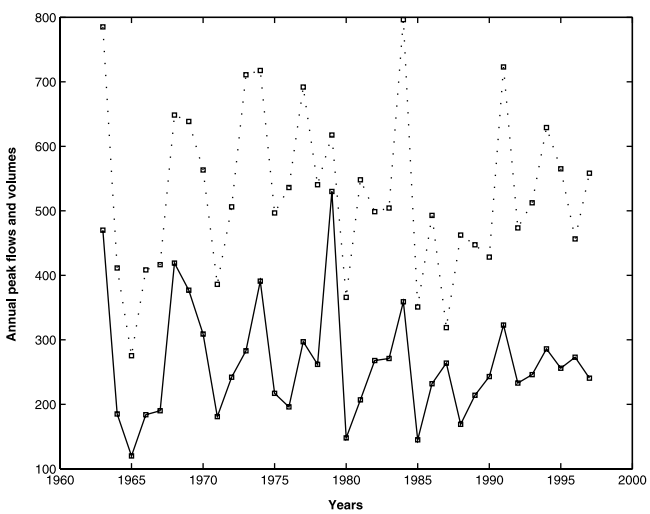

Figure 14. Time series of annual flow and volumes. The upper curve represents annual volumes.

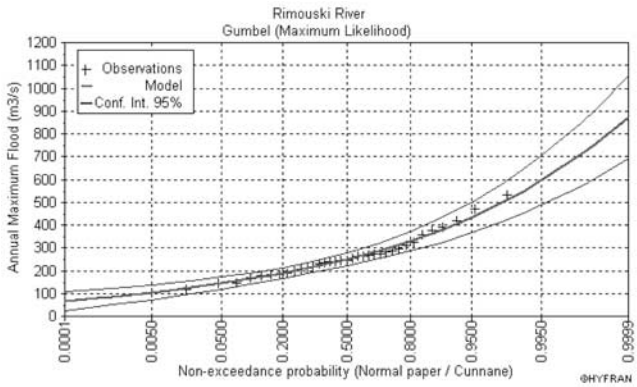

Figure 15. Observed and estimated flows with 95% confidence interval.

[48] In a second step we modeled the link between the two random variables. Note that the correlation coefficients of Spearman and Kendall are respectively of $\rho_S = 0.32$ and $\tau = 0.64$. We considered the same copulas as in section 3 except for Farlie-Gumbel-Morgenstern. The FGM copula can only model a restricted type of dependence: $\tau \in [-2/9, 2/9]$. As stated by *Joe* [1997], this limits the usefulness of this family for modeling purposes. Thus we retained for this application the independence, Clayton and Frank copulas. We estimated the parameters with the same method as in section 3.2. In the case of the Clayton copula, the parameter of dependence was estimated using Kendall's tau coefficient. We obtained $\hat{\alpha}_C = \frac{2\tau}{1-\tau} = 3.55$. For the Frank copula the method of inference functions for margins (IFM) was used, leading to $\hat{\alpha}_F = 9.10$.

4.3. Results

[49] For all considered copulas, a simulation of size 15 000 has been realized using the same technique as in section 3.3. Results are shown in Figure 17.

[50] Again a statistical test has been applied to choose between Frank and Clayton copulas. The coefficients of determination obtained are respectively of 0.97 and 0.96, which shows that Frank copulas seem to have a slightly better behavior. One information of particular interest for hydrologists and water resources managers is the conditional

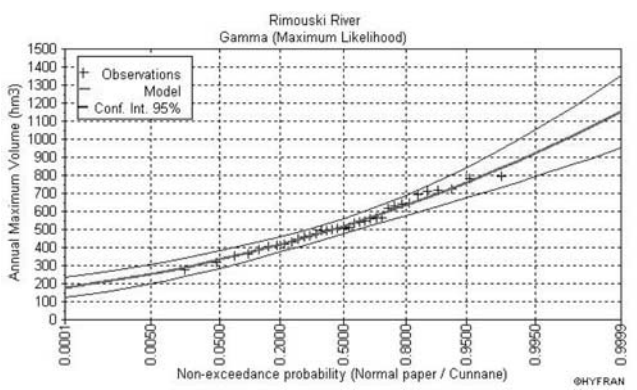

Figure 16. Observed and estimated volumes with 95% confidence interval.

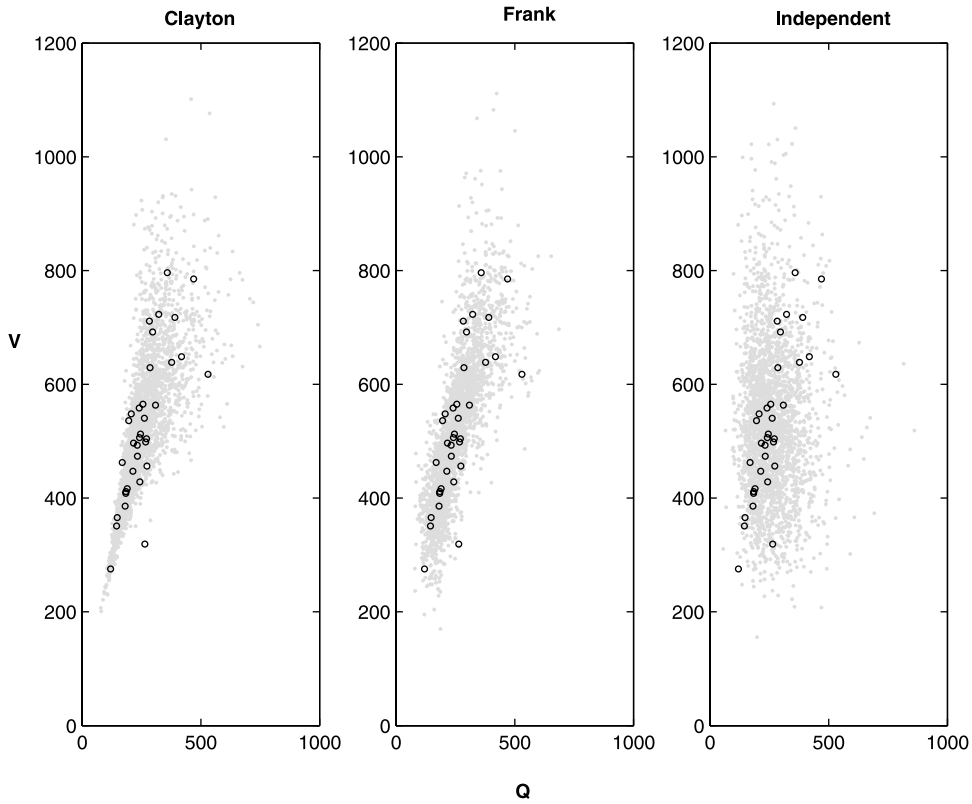

Figure 17. Observations versus simulations for the three types of copulas.

probability of the volume given a specific flow value, say the 100-years return period flow. We are interested in $\Pr(V \leq v | 500 \leq Q \leq 600)$. Figures 18, 19, and 20 illustrate the conditional distribution $\Pr(V \leq v | 500 \leq Q \leq 600)$ for the three models.

[51] Note that the interval [500, 600] corresponds to the 95% asymptotic confidence interval of the 100-years return period flow. Each plot presents the marginal and the conditional distributions. In the case of the independent copula (Figure 18) the difference between the marginal and the conditional distribution should be null if we had considered more simulations. Figures 19 and 20 show that the conditional estimation is more precise since the obtained variance is smaller.

[52] The conditional exceedance probability of a volume, given the value of an extreme flood (100 or 1000 year return period for example), is of major importance for the management of a reservoir. Comparing the conditional density

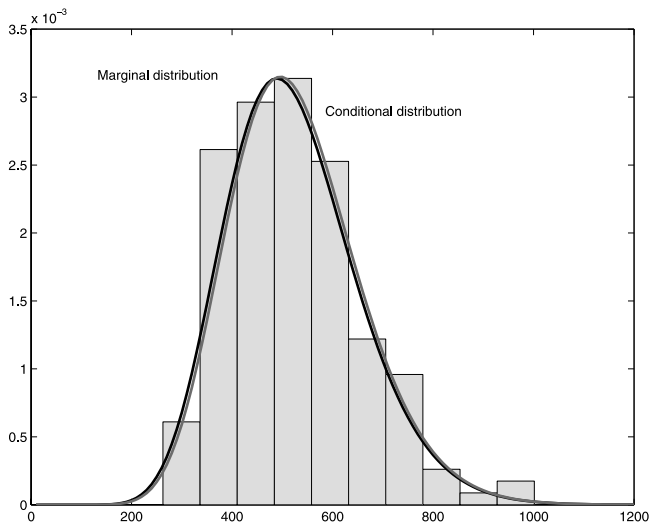

Figure 18. Conditional distribution of the volume given the flow for the independent case.

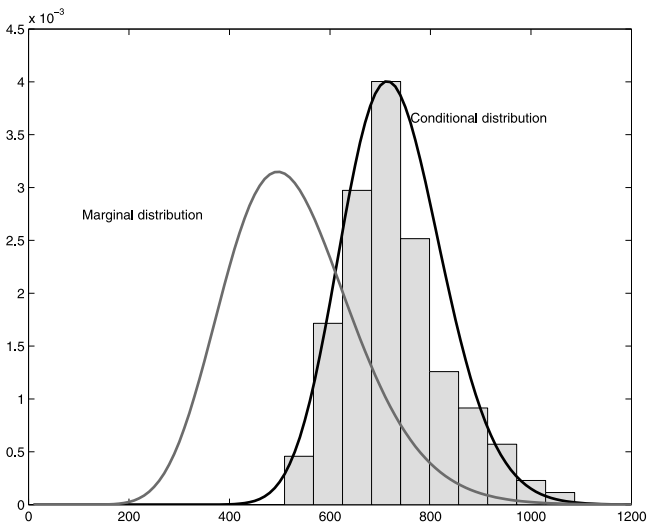

Figure 19. Conditional distribution of the volume given the flow for Frank copula.

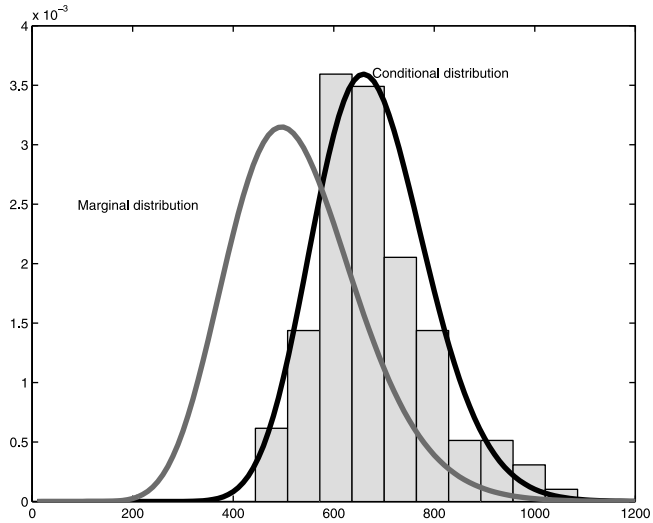

Figure 20. Conditional distribution of the volume given the flow for Clayton copula.

obtained for the independent case with the ones obtained with the two other copulas, it is clear that such exceedance probability would differ substantially. The decisions stemming from this analysis would therefore be different.

5. Conclusions

[53] In this paper we proposed an approach based on copulas applied to bivariate frequency analysis. To our knowledge, such an approach has not been used in hydrology. The model was applied to two different problems in hydrology: flow combination and joint modeling of flow and volume. In the first case the measures of correlation are low and even in this case the difference between the independence case and the Frank copula is of 5.5%. If the correlation would be higher we believe that the difference would be significantly increased. Differences between the copulas should increase as well. The second application emphasizes on the conditional return probability of the flow given the volume. The obtained bivariate probabilities are more precise, which is a serious gain for water resources managers.

[54] The present approach using copulas is promising since it allows to take into account a wide range of correlation, frequently observed in hydrology. In fact the classical multivariate models can not reproduce all type of correlations. Moreover, the standard models are limited, especially because the choice of the marginal distributions is restricted.

[55] The crucial step in the modeling process is the choice of the copula function, which best fits the data. Further work is needed to choose the best copulas able to reproduce the dependence structure of multivariate hydrological variables. We are considering the use of copulas with two or more parameters, for example the Archimax copulas [*Capéraà et al.*, 2000], which encompass the Archimedean copulas and extreme value distributions as special cases. We also propose to estimate the parameters using a Bayesian approach. This method is more suitable than the maximum likelihood when the sample size is small as it is usually the case in hydrology. In the prior distribution we could for example relate Kendall's tau or Spearman's rho with physiographic data (like watershed area, slope. . .), which are always available in practical cases. The trivariate modeling of flow, volume and duration is also of great interest for hydrologists.

[56] **Acknowledgments.** The authors acknowledge the financial support of Hydro-Qubec for this project. We are also grateful to Alcan for the use of the data at Chute-des-Passes.

References

Adamson, P. T., A. V. Metcalfe, and B. Parmentier (1999), Bivariate extreme value distributions: An application of the Gibbs sampler to the analysis of floods, *Water Resour. Res.*, *35*, 2825–2832.

Ali, M. M., N. N. Mikhail, and M. S. Haq (1978), A class of bivariate distributions including the bivariate logistic, *J. Multivariate Anal.*, *8*, 405–412.

Anderson, T. W. (1958), *An Itroduction to Multivariate Statistical Analysis*, John Wiley, Hoboken, N. J.

Bagdonavicius, V., S. Malov, and M. Nikulin (1999), Characterizations and parametric regression estimation in Archimedian copulas, *J. Appl. Stat. Sci.*, *89*, 137–154.

Barbe, P., C. Genest, K. Ghoudi, and B. Rémillard (1996), On Kendall's process, *J. Multivariate Anal.*, *58*, 197–229.

Blest, D. C. (2000), Rank correlation—An alternative measure, *Aust. N. Z. J. Stat.*, *42*, 101–111.

Bouyé, E., A. Durrleman, A. Nikeghbali, G. Riboulet, and T. Roncalli (2000), Copulas for finance—A reading guide and some applications, technical report, Groupe de Rech. Opér., Crédit Lyonnais, Paris.

Bruneau, P., F. Ashkar, and B. Bobée (1994), Simplnorm: Un modèle simple pour obtenir les probabilités conjointes de deux débits et le niveau qui en dépend, *Can. J. Civ. Eng.*, *5*, 883–895.

Capéraà, P., A. L. Fougères, and C. Genest (2000), Bivariate distributions with given extreme value attractor, *J. Multivariate Anal.*, *72*, 30–49.

Clayton, D. G. (1978), A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, *65*, 141–151.

Cook, D. R., and M. E. Johnson (1981), A family of distributions for modeling non-elliptically symmetric multivariate data, *J. R. Stat. Soc. London, Ser. B*, *43*, 210–218.

Cuadras, C. M., and J. Augé (1981), A continuous general multivariate distribution and its properties, *Commun. Stat. A Theory Methods*, *10*, 339–353.

Deheuvels, P. (1979), La fonction de dépendance empirique et ses propriétés—Un test non paramétrique d'indépendance, Bull, *Cl. Sci. Acad. R. Belg., 5 Ser.*, *65*, 274–292.

De Michele, C., and G. Salvadori (2003), A generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas, *J. Geophys. Res.*, *108*(D2), 4067, doi:10.1029/2002JD002534.

Embrechts, P., A. McNeil, and D. Straumann (2002), Correlation and dependence in risk management: Properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, pp. 176–223, Cambridge Univ. Press, New York.

Embrechts, P., F. Lindskog, and A. McNeil (2003), Modelling dependence with copulas and applications to risk management, in *Handbook of Heavy Tailed Distributions in Finance*, pp. 329–384, Elsevier Sci., New York.

Eyraud, H. (1938), Les principes de la mesure des corrélations, *Ann. Univ. Lyon, Ser. A*, *1*, 30–47.

Farlie, D. G. J. (1960), The performance of some correlation coefficients for a general bivariate distribution, *Biometrika*, *47*, 307–323.

Favre, A.-C., A. Musy, and S. Morgenthaler (2002), Two-site modeling of rainfall based on the Neyman-Scott process, *Water Resour. Res.*, *38*(12), 1307, doi:10.1029/2002WR001343.

Frank, M. J. (1979), On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$, *Aequationes Math.*, *19*, 194–226.

Frees, E. W., and E. A. Valdez (1998), Understanding relationships using copulas, *N. Am. Actuarial J.*, *2*(1), 1–25.

Galambos, J. (1975), Order statistics of samples from multivariate distributions, *J. Am. Stat. Assoc.*, *70*, 674–680.

Genest, C. (1987), Frank's family of bivariate distributions, *Biometrika*, *74*, 549–555.

Genest, C., and K. Ghoudi (1994), Une famille de lois bidimensionnelles insolite, *C. R. Acad. Sci. Paris, Ser. I*, *318*, 351–354.

Genest, C., and R. J. MacKay (1986a), Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données, *Can. J. Stat.*, *14*, 145–159.

Genest, C., and R. J. MacKay (1986b), The joy of copulas: Bivariate distributions with uniform marginals, *Am. Stat.*, *40*, 280–283.

Genest, C., and J.-F. Plante (2003), On Blest's measure of rank correlation, *Can. J. Stat.*, *31*(1), 1–18.

Genest, C., and L.-P. Rivest (1993), Statistical inference procedure for bivariate Archimedean copulas, *J. Am. Stat. Assoc.*, *88*, 1034–1043.

Gumbel, E. J. (1960), Bivariate exponential distributions, *J. Am. Stat. Assoc.*, *55*, 698–707.

Gumbel, J. (1958), *Statistics of Extremes*, Columbia Univ. Press, New York.

Hougaard, P. (1986), A class of multivariate failure time distributions, *Biometrika*, *73*, 671–678.

Hüsler, J., and R. D. Reiss (1989), Maxima of normal random vectors: Between independence and complete dependence, *Stat. Prob. Lett.*, *7*, 283–286.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, New York.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*, John Wiley, Hoboken, N. J.

Johnson, N., and S. Kotz (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley, Hoboken, N. J.

Johnson, N., S. Kotz, and N. Balakrishnan (1997), *Discrete Multivariate Distributions*, John Wiley, Hoboken, N. J.

Johnson, R., and D. Wichern (1988), *Applied Multivariate Statistical Analysis*, Prentice-Hall, Old Tappan, N. J.

Jorgensen, B. (1997), *The theory of Dispersion Models*, Chapman and Hall, New York.

Krzanowski, W. J. (1988), *Principles of Multivariate Analysis: A User's Perspective*, Oxford Univ. Press, New York.

Lee, A. J. (1993), Generating random binary deviates having fixed marginal distributions and specified degrees of association, *Am. Stat.*, *47*, 209–215.

Long, D., and R. Krzysztofowicz (1992), Farlie-Gumbel-Morgenstern bivariate densities: Are they applicable in hydrology?, *Stochastic Hydrol. Hydraul.*, *6*, 47–54.

Morgenstern, D. (1956), Einfache Beispiele Zweidimensionaler Verteilungen, *Mitt. Math. Stat.*, *8*, 234–235.

Nelsen, R. B. (1986), Properties of a one-parameter family of bivariate distribution with specified marginals, *Commun. Stat. Theory Methods*, *15*, 3277–3285.

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

Oakes, D. (1982), A model fo association in bivariate survival data, *J. R. Stat. Soc. London, Ser. B*, *44*, 414–422.

Perreault, L. (2003), Modélisation des débits de pointe sortants Chute-des-Passes: Application du mélange de distributions normales, *Tech. Rep. IREQ-2003-127C*, Inst. de Rech. d'Hydro-Québec, Varennes, Quebec, Canada.

Raynal-Villasenor, J. A., and J. D. Salas (1987), Multivariate extreme value distributions in hydrological analyses, in *Water for the Future: Hydrology in Perspective*, p. 111–119, Int. Assoc. of Hydrol. Sci., Gentbruge.

Robert, C. P. (1996), Inference in mixture models, in *Markov Chain Monte Carlo in Practice*, pp. 441–464, Chapman and Hall, New York.

Schweizer, B. (1991), Thirty years of copulas, in *Advances in Probability Distributions with Given Marginals*, pp. 13–50, Kluwer Acad., Norwell, Mass.

Schweizer, B., and E. F. Wolff (1981), On nonparametric measures of dependence for random variables, *Ann. Stat.*, *9*, 879–885.

Singh, K., and V. P. Singh (1991), Derivation of bivariate probability density functions with exponential marginals, *Stochastic Hydrol. Hydraul.*, *5*, 55–68.

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

Song, P. X. K. (2000), Multivariate dispersion models generated from Gaussian copulas, *Scand. J. Stat.*, *27*, 305–320.

Stedinger, J. R., R. M. Vogel. and E. Foufoula-Georgiou (1993), Frequency analysis of extreme events, in *Handbook of Hydrology*, pp. 18.1–18.66, McGraw-Hill, New York.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley, Hoboken, N. J.

Wang, W. J. (2001), A Bayesian joint probability approach for flood record augmentation, *Water Resour. Res.*, *37*, 1707–1712.

West, M. (1992), Modelling with mixtures, in *Bayesian Statistics 4*, pp. 503–525, Oxford Univ. Press, New York.

Yue, S., T. B. M. J. Ouarda, and B. Bobée (2001), A review of bivariate gamma distribution for hydrological application, *J. Hydrol.*, *246*, 1–18.

B. Bobée, S. El Adlouni, and A.-C. Favre, Chaire en hydrologie statistique, INRS, Eau-Terre et Environnement, Université du Québec, 2800 rue Einstein, Sainte-Foy, Québec, Canada G1V 4C7. (anne-catherine_favre@inrs-ete.uquebec.ca)

L. Perreault, Institut de Recherche d'Hydro-Québec, Varennes, Québec, Canada J3X 1S1.

N. Thiémonge, Conception des aménagements de production, Hydro-Québec, Montréal, Québec, Canada H2L 4P5.