Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data

C. Genest,[1] A.-C. Favre,[2] J. Béliveau,[3] and C. Jacques[4]

Received 23 June 2006; revised 9 February 2007; accepted 29 May 2007; published 5 September 2007.

[1] Metaelliptical copulas are introduced as a flexible tool for modeling multivariate data in hydrology. The properties of this broad class of dependence functions are reviewed, along with associated rank-based procedures for copula parameter estimation and goodness-of-fit testing. A new graphical diagnostic tool is also proposed for selecting an appropriate metaelliptical copula. Peak, volume, and duration of the annual spring flood for the Romaine River (Québec, Canada) are used for illustration purposes.

Citation: Genest, C., A.-C. Favre, J. Béliveau, and C. Jacques (2007), Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data, *Water Resour. Res.*, *43*, W09401, doi:10.1029/2006WR005275.

1. Introduction

[2] Copula modeling is quickly gaining in popularity for the treatment of multivariate hydrologic data. Starting with the papers of *De Michele and Salvadori* [2003] and *Favre et al.* [2004], this approach has been used profitably, e.g., in the study of storm rainfall characteristics and in flood frequency analysis. See *Salvadori and De Michele* [2004, 2006], *De Michele et al.* [2005], *Zhang and Singh* [2006], or *Grimaldi and Serinaldi* [2006] for additional examples of applications, as well as *Genest and Favre* [2007] for a review of inference techniques for copula modeling in a hydrologic setting.

[3] Given a vector $\mathbf{X} = (X_1,\ldots,X_p)$ of $p \geq 2$ continuous random variables, the copula approach hinges on the representation

$$\Pr\left(X_1 \leq x_1,\ldots,X_p \leq x_p\right) = C\left\{F_1(x_1),\ldots,F_p\left(x_p\right)\right\} \qquad (1)$$

for the joint distribution of $\mathbf{X}$ in terms of its marginal distributions

$$F_k(x) = \Pr(X_k \leq x), \qquad k \in \{1,\ldots,p\}$$

and a copula C, i.e., the cumulative distribution function of a vector $(U_1,\ldots,U_p)$ of dependent uniform random variables on the interval (0, 1).

[4] From a theoretical point of view, copulas are attractive because of the flexibility they offer in the construction of models for the vector $\mathbf{X}$ through the choice of margins from different families of univariate distributions and, quite separately, the selection of a suitable dependence structure between the components of $\mathbf{X}$, as represented by C.

[5] From a practical point of view, the interest in this methodology stems from the observation that most hydrologic phenomena are multifactorial and that dependence between variables must be accounted for in order to achieve realistic modeling. For example, hydrologic engineers face planning, design and management problems that require a detailed knowledge of the three main flood characteristics: peak, volume and duration. Such is the case, e.g., for flooding and inundation management.

[6] Unfortunately, most frequency analysis applications considered to date have focused either on one or two variables at the time. As a result, they cannot provide a complete assessment of the probability of flood occurrence. As shown by *De Michele et al.* [2005], among others, failure to take into account the dependence between all relevant variables may lead to an overestimation or underestimation of the risk associated with a given event.

[7] To this date, copula modeling involving more than two hydrologic variables has only been attempted by *Salvadori and De Michele* [2006] and *Grimaldi and Serinaldi* [2006] in the context of storm and flood analysis, respectively. In the latter paper, the authors consider the use of Archimedean copulas for joint modeling of flood peak, volume and duration. These copulas are expressible in the form

$$C\left(u_1,\ldots,u_p\right) = \phi^{-1}\left\{\phi(u_1) + \cdots + \phi\left(u_p\right)\right\} \qquad (2)$$

in terms of a generator ϕ: (0, 1] $\rightarrow$ [0, ∞) such that $\phi(1) = 0$ and other regularity conditions are satisfied, e.g., $(-1)^k d\phi^{-1}(t)/dt^k \geq 0$ for every $k \in \{1,\ldots,p\}$.

[8] Many examples of Archimedean copulas are given, e.g., in chapter 4 of *Nelsen* [2006]. Although this class is broad, *Grimaldi and Serinaldi* [2006] argue that it is often too restrictive for hydrologic applications. For, the symmetry of (2) implies that all pairs of variables share the same dependence structure and hence the same degree of association as measured by margin-free coefficients such as Spearman's rho or Kendall's tau.

[1] Département de mathématiques et de statistique, Université Laval, Quebec, Quebec, Canada.
[2] Chaire en hydrologie statistique, Centre Eau, Terre et Environnement, Institut National de la Recherche Scientifique, Quebec, Quebec, Canada.
[3] Direction des politiques et programmes, Ministère de l'Éducation, du Loisir et du Sport, Quebec, Quebec, Canada.
[4] Service du développement et des infrastructures stratégiques, Ministère des Affaires Municipales et des Régions, Quebec, Quebec, Canada.

0043-1397/07/2006WR005275$09.00

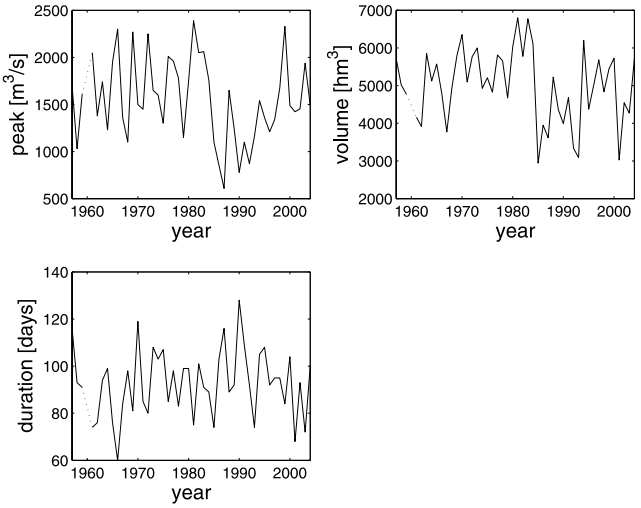

Figure 1. Time series plot of peak (m^3/s), volume (hm^3), and duration (days) for the Romaine River.

[9] Following *Joe* [1997], *Grimaldi and Serinaldi* [2006] are thus led to consider nested classes of Archimedean copulas, which may involve up to $p - 1$ generators ϕ_1, ..., ϕ_{p-1} that must then satisfy a wealth of regularity and compatibility conditions. However, even in the simplest example of this construction, namely,

$$C(u_1, u_2, u_3) = \phi_1^{-1}\left[\phi_1(u_3) + \phi_1 \circ \phi_2^{-1}\{\phi_2(u_2) + \phi_2(u_1)\}\right]; \quad (3)$$

some pairs still have the same dependence structure, e.g., (U_1, U_3) and (U_2, U_3).

[10] The purpose of this paper is to show how greater flexibility can be achieved in modeling trivariate hydrologic data using metaelliptical families of copulas. These models are not new: they were originally proposed by *Fang et al.* [2002] and they are already beginning to find applications in finance; see, e.g., *Cherubini et al.* [2004, and references therein]. However, their properties are scattered in the literature and inference procedures for this class of copulas are still in their early stages of development.

[11] In this paper, metaelliptical copulas are used to analyze flood peak, volume and duration data for the Romaine River, located in the Basse-Côte-Nord area of Québec (Canada). Beyond its intrinsic interest for hydrologists, this application is only one of a handful applications (finance included) to illustrate the merits of metaelliptical copulas for dependence modeling in a truly multivariate (as opposed to bivariate) context. This paper is also among the first to implement the goodness-of-fit tests of *Genest et al.* [2006] and *Genest and Rémillard* [2007] to this important class of copulas.

[12] The hydrologic data at the origin of this work are presented in section 2. A review of the definition and basic properties of metaelliptical copulas is then given in section 3, along with a compendium of the most common parametric families of this form. As mentioned earlier, this material is gathered from scattered sources in the recent statistical and financial literature; a unified treatment is presented here with the end user in mind.

[13] General estimation and goodness-of-fit procedures for copula models are adapted to the metaelliptical class in section 4. To help analysts choose from a variety of models from this class, an inference procedure is required for selecting their generator. Estimation and graphical diagnostic tools of this sort are presented here for the first time; they rely on nonparametric techniques that prevail in copula modeling methodology. These inference techniques are then used in section 5 to analyze the Romaine River annual spring flood data. Concluding comments may be found in section 6.

2. Motivation: Data

[14] Peak (m^3/s), volume (hm^3) and duration (days) of the annual spring flood are available from 1957 to 2004 for the Romaine River located in the Basse-Côte-Nord area of Québec (Canada). The series consist of $n = 47$ observations, as the data for 1960 are missing.

[15] Figure 1 shows the evolution of the three variables during the study period. As will be shown in section 5.1, the individual time series are stationary and exhibit no autocorrelation; accordingly, they can be assimilated to random samples from univariate distributions whose form will be determined later. A careful look at Figure 1 suggests that while the variables may not depend on time, they are related to one another. To confirm this suspicion, scatterplots of each pair of variables could be drawn. However, as argued by *Genest and Favre* [2007], among others, dependence between variables is best revealed by plotting the ranks of the data, rather than the original variables. This is done in Figure 2 for each of the pairs (peak, volume), (peak, duration), and (volume, duration).

[16] The rank plots in Figure 2 clearly suggest the presence of positive dependence in the pair (peak, volume) and negative dependence in the pair (peak, duration). This conforms to intuition. There is also a hint of positive association in the pair (volume, duration).

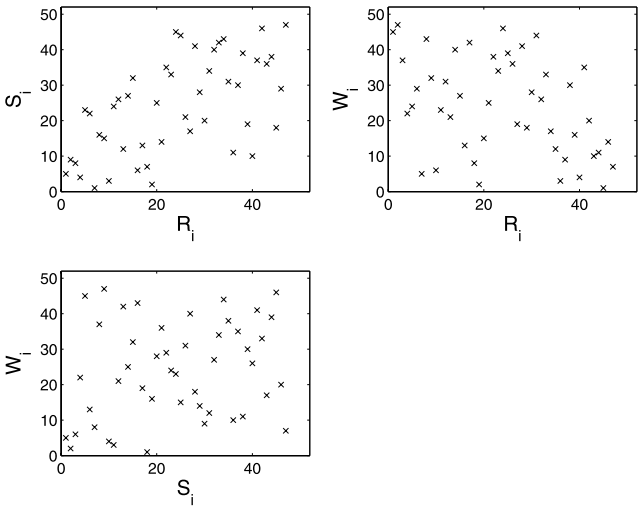

Figure 2. Scatterplots of pairs of ranks for peak (R) versus volume (S), peak (R) versus duration (W), and volume (S) versus duration (W) for the Romaine River, Québec (Canada), based on annual data collected between 1957 and 2004.

Table 1. Dependence Measures for the Pairs of the Romaine River, Québec, Canada, Based on Annual Data Collected Between 1957 and 2004[a]

	Peak and Volume	Peak and Duration	Volume and Duration
r_n	0.5832 [<0.0001]	−0.4756 [0.0007]	0.2124 [0.1516]
ρ_n	0.5892 [<0.0001]	−0.3943 [0.0075]	0.2259 [0.1255]
τ_n	0.4070 [<0.0001]	−0.2812 [0.0058]	0.1471 [0.1491]

[a]*P* values are given in brackets.

[17] To confirm these findings, one might compute Pearson's correlation, r_n, and base a test of independence thereupon. Although the results are reported in Table 1 for the three pairs of variables, the conclusions are somewhat dubious because they are based on the assumption of bivariate normality, which turns out to be inappropriate in this case.

[18] A better way to proceed consists of computing Spearman's correlation, ρ_n, or Kendall's coefficient of concordance, τ_n, and to base a test of independence on either one of these statistics. Because they are rank based, these procedures are known to be robust to departures from normality, while remaining powerful; see, e.g., *Genest and Verret* [2005] for a recent discussion.

[19] Empirical values of ρ_n and τ_n are also reported in Table 1, along with the *P* values of the corresponding tests of independence. Although volume, peak and duration may be construed as continuous attributes of the annual spring flood, some tied observations occurred among the latter two because of the coarseness in the scales on which the data were recorded. At the cost of a small bias, the estimates in Table 1 are thus based on Spearman's ρ_s and Kendall's τ_b coefficients as defined, e.g., by *Kendall* [1955].

[20] Archimedean copulas (2) are clearly inadequate for modeling the dependence between the three variables, given that different pairs exhibit wildly varying degrees of dependence. Copulas of the form (3) would not be of any help either, because the regularity conditions imposed by *Joe* [1997] and *Grimaldi and Serinaldi* [2006] imply that negative degrees of association cannot be modeled between the variables [see, e.g., *Marshall and Olkin*, 1988].

[21] While it may be argued that the dependence in the pair (volume, duration) is not significant, this does not imply that they are independent. (Recall that zero correlation and independence are not equivalent, except under normality.) Furthermore, the option of treating, say, duration as independent of the other pair (peak, volume) is not viable, because of the demonstrated dependence between peak and duration. Trivariate modeling thus imposes itself in this case.

3. Metaelliptical Copulas

[22] The class of metaelliptical copulas was originally introduced by *Fang et al.* [2002]. It is derived from the well-known family of elliptical distributions, which is itself an extension of the classical multivariate normal distribution. For a detailed overview on elliptical distributions [see, e.g., *Fang et al.*, 1990].

[23] Specifically, a p variate vector $\mathbf{X}^* = (X_1^*,\ldots,X_p^*)$ is said to have an elliptical distribution $\mathcal{E}_p(\mu, \Sigma, g)$ with mean vector $\mu \in \mathbb{R}^p$, covariance matrix $\Sigma = (\sigma_{ij})$ and generator g: $[0, \infty) \to [0, \infty)$ if it can be expressed in the form

$$\mathbf{X}^* = \mu + RA\mathcal{U}, \tag{4}$$

where $AA^\intercal = \Sigma$ is the Cholesky decomposition of Σ, $\mathcal{U}$ is a p variate random vector uniformly distributed on the sphere $\mathbb{S}_p = \{(u_1,\ldots,u_p) \in \mathbb{R}^p: u_1^2 + \cdots + u_p^2 = 1\}$ and R is a nonnegative random variable with density

$$f_g(r) = \frac{2\pi^{p/2}}{\Gamma(p/2)}\, r^{p-1} g\left(r^2\right), \qquad r > 0.$$

The representation (4) is such that when it exists, the multivariate density of the vector $\mathbf{X}^*$ is given by $h_g(\mathbf{x}) = |\Sigma|^{-1/2} g\{(\mathbf{x} - \mu)^\intercal \Sigma^{-1} (\mathbf{x} - \mu)\}$.

[24] When $g(t) \propto e^{-t/2}$, for instance, $\mathbf{X}^*$ is multivariate normal, and R^2 has a chi-square distribution with p degrees of freedom. Other common examples of g are given in Table 2; they lead, e.g., to the multivariate Student or Pearson type II distributions.

[25] Through an appropriate choice of parameters, the generators listed in Table 2 provide added flexibility in modeling p variate data. In particular, they allow for fatter tails than under normality and, more importantly, for the possibility of tail dependence [*Joe*, 1997].

[26] Tail dependence, say between components k and ℓ of $\mathbf{X}^*$, can be represented by the probability that X_k^* exceeds its quantile of order α, given that X_ℓ^* exceeds its own quantile of order α. The limiting probability as $\alpha \to 1$ is called the upper tail dependence coefficient. When the vector $\mathbf{X}^*$ is elliptically distributed, *Schmidt* [2002] shows that

$$\lambda\left(X_k^*, X_\ell^*\right) = \int_0^{s_{k\ell}} \frac{t^\beta}{\sqrt{1-t^2}}\, dt \Big/ \int_0^1 \frac{t^\beta}{\sqrt{1-t^2}}\, dt,$$

Table 2. Useful Functions for the Simulation of Four Metaelliptical Copulas

Copula	Distribution of R^2	$g(t)$	Q_g
Normal	$R^2 \sim \chi^2_{(p)}$	$(2\pi)^{-p/2} \exp(-t/2)$	$\mathcal{N}(0, 1)$
Student	$R^2/p \sim \mathcal{F}(p, \nu)$	$\frac{(\pi\nu)^{-p/2}\Gamma\left(\frac{p+\nu}{2}\right)}{\Gamma(\nu/2)}(1 + t/\nu)^{-(p+\nu)/2}$	Student (ν)
Cauchy	$R^2/p \sim \mathcal{F}(p, 1)$	$\frac{(\pi)^{-p/2}\Gamma\left(\frac{p+1}{2}\right)}{\Gamma(1/2)}(1 + t)^{-(p+1)/2}$	Cauchy
Pearson type II	$R^2 \sim \text{Beta}(p/2, \nu + 1)$	$\frac{\Gamma(p/2+\nu+1)}{\pi^{p/2}\Gamma(\nu+1)}(1 - t)^\nu$, $t \in [-1, 1]$, $\nu > -1$	Pearson type II

where $s_{k\ell} = \sqrt{(1+\rho_{k\ell})/2}$ with

$$\rho_{k\ell} = \frac{\sigma_{k\ell}}{\sqrt{\sigma_{kk}\sigma_{\ell\ell}}}, \qquad k,\ell \in \{1,\ldots,p\}$$

and $\beta > 0$ is such that for arbitrary $x > 0$, $g(xt)/g(t) \to x^{-(p+\beta)/2}$ as $t \to \infty$.

[27] In this case as in general, the upper tail coefficient $\lambda \in [0, 1]$ is independent of the marginal distributions. As illustrated by *Frahm et al.* [2003], among others, it is nonzero for many metaelliptical copula models; a contrario, it vanishes for the multivariate normal distribution unless $\rho_{k\ell} = 1$ for all $k, \ell \in \{1,\ldots,p\}$. Refer to *Poulin et al.* [2007] for a discussion of tail dependence and its importance in the context of bivariate frequency analysis.

[28] One inconvenient limitation of elliptical distributions is that the scaled components $X_1^*/\sqrt{\sigma_{11}},\ldots,X_p^*/\sqrt{\sigma_{pp}}$ are identically distributed. Thus for every $k \in \{1,\ldots,p\}$,

$$Q_g(x) = \Pr\left(\frac{X_k^*}{\sqrt{\sigma_{kk}}} \le x\right) = \frac{1}{2} + \frac{\pi^{(p-1)/2}}{\Gamma\left(\frac{p-1}{2}\right)} \int_0^x \int_{u^2}^\infty \left(y-u^2\right)^{(p-1)/2-1} g(y)\,dy\,du. \quad (5)$$

However, models based on the unique metaelliptical copula $C_{\Sigma,g}$ associated with $\mathbf{X}^*$ do not suffer from this defect, as their margins are arbitrary.

[29] Formally $C_{\Sigma,g}$ is the joint distribution of the vector $(\mathcal{U}_1,\ldots,\mathcal{U}_p)$ with $\mathcal{U}_k = Q_g(X_k^*/\sqrt{\sigma_{kk}})$ for $k \in \{1,\ldots,p\}$. The cumulative distribution function of an elliptical vector $\mathbf{X}^*$ (and hence of its components) is typically not available in closed form. Consequently, its inverse is not explicit and an expression for $C_{\Sigma,g}(u_1,\ldots,u_p)$ is neither useful nor enlightening.

[30] A better understanding of the structure of elliptical distributions can be derived from representation (4) which justifies the following random generation algorithm.

[31] **Algorithm 1**. To generate a p variate observation $\mathbf{X}^*$ from $\mathcal{E}_p(\mu, \Sigma, g)$, proceed as follows:

[32] 1. Generate R according to distribution f_g.

[33] 2. Generate $\mathcal{U}$ uniformly on S_p, as per, e.g., *Marsaglia* [1972].

[34] 3. Compute the square root A of Σ via the Cholesky decomposition.

[35] 4. Deliver $\mathbf{X}^* = \mu + RA\mathcal{U}$.

[36] To see what this does, consider the special case where $\mu = 0$ and $\Sigma = I_p$ is the identity matrix. An observation of $\mathbf{X}^*$ is then obtained by determining at what distance R it will be from the origin, and in which direction. This is what steps 1 and 2 of the algorithm do. Given that all directions are equally likely, the resulting distribution is spherical, meaning that its lines of isodensity are concentric circles centered at 0 in $\mathbb{R}^p$.

[37] For example, it was mentioned earlier that when $g(t) \propto e^{-t/2}$, R^2 is chi-square with p degrees of freedom and that as a consequence, $R\mathcal{U}$ has a standard p variate normal distribution. The introduction of $A = \Sigma^{1/2}$ in step 3 of the algorithm makes it possible to transform these concentric circles into ellipses, and $AR\mathcal{U}$ is then multivariate normal with covariance matrix Σ. Step 4 then moves the center of the distribution to μ.

[38] Now suppose that an observation $\mathbf{X} = (X_1,\ldots,X_p)$ from a copula model of the form (1) is desired, where C is metaelliptical with functional parameter g. The procedure then consists of generating $\mathbf{X}^* = (X_1^*,\ldots,X_p^*)$ according to algorithm 1, after which the margins are transformed to $F_1,\ldots,F_p$ by setting $X_k = F_k^{-1} \circ Q_g(X_k^*/\sqrt{\sigma_{kk}})$ for all $k \in \{1,\ldots,p\}$. Given that the effect of translation and scaling of the individual components is canceled in this operation, no loss in generality incurs from taking $\mu = 0$ and $\sigma_{11} = \cdots = \sigma_{pp} = 1$. Consequently, Σ is assumed to represent a correlation matrix in the sequel.

[39] **Algorithm 2**. To generate a p variate observation $\mathbf{X}$ from copula $C_{\Sigma,g}$, proceed as follows:

[40] 1. Carry out steps 1–4 of algorithm 1 resulting in a p variate vector $\mathbf{X}^* = (X_1^*,\ldots,X_p^*)$.

[41] 2. Deliver $\mathbf{X} = (X_1,\ldots,X_p)$, where $X_k = F_k^{-1} \circ Q_g(X_k^*/\sqrt{\sigma_{kk}})$ for all $k \in \{1,\ldots,p\}$.

[42] The main properties of metaelliptical copulas are described by *Fang et al.* [2002] and *Abdous et al.* [2005], among others. Of particular importance here is a result of *Hult and Lindskog* [2002] to the effect that for every $k, \ell \in \{1,\ldots,p\}$,

$$\mathrm{corr}\left(X_k^*, X_\ell^*\right) = \rho_{k\ell} = \frac{\sigma_{k\ell}}{\sqrt{\sigma_{kk}\sigma_{\ell\ell}}}$$

is linked to the population value of Kendall's tau between X_k^* and X_ℓ^* through the relation

$$\tau_{k\ell} = \tau\left(X_k^*, X_\ell^*\right) = \frac{2}{\pi}\arcsin(\rho_{k\ell}), \qquad k,\ell \in \{1,\ldots,p\}. \quad (6)$$

[43] Given that Kendall's tau is invariant by monotone increasing functions of the margins, equation (6) also describes the connection between $\rho_{k\ell}$ and $\tau(X_k, X_\ell) = \tau_{k\ell}$ for every choice of $k, \ell \in \{1,\ldots,p\}$. One point worth stressing is the fact that unless g is the generator of the multivariate normal distribution, $\tau_{k\ell} = \rho_{k\ell} = 0$ never corresponds to independence, either between X_k and X_ℓ or between X_k^* and X_ℓ^*.

4. Inference Procedures

[44] In order to model the dependence between $p \ge 2$ random variables using a metaelliptical copula, appropriate choices of Σ and g must be made. Estimation methods for both of these parameters are discussed in turn.

[45] Given that the copula of a random vector is unaffected by monotone increasing transformations of the margins, it is taken for granted here that copula inference should be based on the maximally invariant statistic of the initial random sample $\mathbf{X}_1 = (X_{11},\ldots,X_{1p}),\ldots,\mathbf{X}_n = (X_{n1},\ldots,X_{np})$. As explained, e.g., by *Genest and Favre* [2007], this implies that estimators and tests for copula structures should be functions of the set of rank vectors $\mathbf{R}_1 = (R_{11},\ldots,R_{1p}),\ldots,\mathbf{R}_n = (R_{n1},\ldots,R_{np})$, where for each $i \in \{1,\ldots,n\}$ and $k \in \{1,\ldots,p\}$,

$$R_{ik} = \sum_{j=1}^n \mathbf{1}\left(X_{jk} \le X_{ik}\right).$$

4.1. Estimation of Σ

[46] Because Σ is a correlation matrix, it is symmetric and hence one need only estimate the vector θ of its $p(p-1)/2$

supradiagonal elements. Given a choice of g, the most efficient way to proceed consists of maximizing the log-pseudo likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log\left[c_\theta\left\{\hat{F}_1(X_{i1}), \ldots, \hat{F}_p(X_{ip})\right\}\right] = \sum_{i=1}^{n} \log\left\{c_\theta\left(\frac{R_{i1}}{n+1}, \ldots, \frac{R_{ip}}{n+1}\right)\right\}.$$

[47] Here, c_θ is the density corresponding to the meta-elliptical copula

$$C_\theta(u_1, \ldots, u_p) = \int_{-\infty}^{Q_g^{-1}(u_1)} \cdots \int_{-\infty}^{Q_g^{-1}(u_p)} \frac{1}{|\Sigma|^{1/2}}\, g\left(\mathbf{z}^\top \Sigma^{-1} \mathbf{z}\right) dz_p \cdots dz_1, \quad (7)$$

where $\mathbf{z} = (z_1, \ldots, z_p)^\top$ and Q_g^{-1} denotes the inverse of Q_g defined in (5). Furthermore,

$$\hat{F}_k(t) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}(X_{ik} \le t)$$

is the rescaled empirical distribution function associated with variable X_k, for $k \in \{1, \ldots, p\}$.

[48] The division by $n + 1$ rather than n in the definition of $\hat{F}_k$ is standard in the copula modeling literature. It ensures that $\hat{F}_k(X_{ik}) = R_{ik}/(n+1) < 1$ for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, p\}$. As a consequence, each summand in $\ell(\theta)$ is finite, which might not be the case if the density c_θ were evaluated at a point on the boundary of $[0,1]^p$.

[49] As shown by *Genest et al.* [1995] and *Shih and Louis* [1995], the value of θ that maximizes the rank-based pseudo likelihood is asymptotically normal under weak regularity conditions. It is also semiparametrically asymptotically efficient in specific circumstances [see *Genest and Werker*, 2002].

[50] In the present context, however, it is numerically more convenient to use another rank-based estimator, which is also consistent [see, e.g., *Genest and Rémillard*, 2007]. The latter is based on the inversion of Kendall's tau; it may be construed as a nonparametric analogue of the celebrated method of moments. In the present case, this approach is facilitated by identity (6). It was exploited in the case of metaelliptical copulas, e.g., by *Breymann et al.* [2003] and by *Lindskog et al.* [2003]. Recalling that Σ is taken here to be a correlation matrix, the method consists of setting

$$\hat{\sigma}_{k\ell} = \sin(\pi \hat{\tau}_{k\ell}/2),$$

where for each k, $\ell \in \{1, \ldots, p\}$, $\hat{\tau}_{k\ell}$ is the sample version of Kendall's tau.

[51] Specifically, for given k, $\ell \in \{1, \ldots, p\}$,

$$\hat{\tau}_{k\ell} = \frac{C_{k\ell} - D_{k\ell}}{C_{k\ell} + D_{k\ell}},$$

where $C_{k\ell}$ and $D_{k\ell}$ represent the number of concordant and discordant pairs, respectively. Distinct pairs $(X_{ik}, X_{i\ell})$ and $(X_{jk}, X_{j\ell})$ are said to be concordant whenever $(X_{jk} - X_{ik})(X_{j\ell} - X_{i\ell}) > 0$ or equivalently if $(R_{jk} - R_{ik})(R_{j\ell} - R_{i\ell}) > 0$; they are said to be discordant otherwise. As $\hat{\tau}_{k\ell}$ is a U statistic, it is well known to be asymptotically normal and unbiased. See, e.g., *Genest and Rivest* [1993] for a consistent estimate of its limiting variance. For additional information about U statistics and their distributional properties, see, e.g., *Lee* [1990].

4.2. Estimation of g

[52] The estimation of g is more complex, considering that it is a functional parameter. Indeed, a rigorous approach to this problem has yet to be developed. Financial applications to date have simply treated g as fixed; however, several possible choices of g have often been considered to assess the robustness of the conclusions derived from the model.

[53] An informal graphical tool is proposed below for assisting in the selection of an appropriate generator g. Goodness-of-fit tests for metaelliptical copula models are then described in the following subsection. Validating the choice of g is of paramount importance, given that it determines tail dependence in metaelliptical copula models [*Schmidt*, 2002].

[54] It is well known that if $\mathbf{X}^*$ is a p variate normal vector with mean $\mu = 0$ and covariance matrix Σ, then $(\mathbf{X}^*)^\top \Sigma^{-1} \mathbf{X}^*$ follows a chi-square distribution with p degrees of freedom [see, e.g., *Rao*, 1981]. More generally if $\mathbf{X}^*$ is distributed as an $\mathcal{E}_p(0, \Sigma, g)$, it then follows from representation (4) that

$$\left(\mathbf{X}^*\right)^\top \Sigma^{-1} \mathbf{X}^* = R^2 (AU)^\top \Sigma^{-1} (AU) = R^2 \quad (8)$$

since $A = \Sigma^{1/2}$ and $U^\top U = 1$. Hence if $\mathbf{X}$ is a p variate random vector with metaelliptical copula C with parameters Σ and g and margins $F_1, \ldots, F_p$, equation (8) then holds for the vector $\mathbf{X}^*$ with components $X_k^* = Q_g^{-1} \circ F_k(X_k)$ for $k \in \{1, \ldots, p\}$.

[55] Now suppose independent copies $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X}$ have been observed. Let $\hat{F}_k$ represent the consistent estimator of F_k defined in section 4.1, and set $\mathbf{Y}_i^* = (Y_{i1}^*, \ldots, Y_{ip}^*)$, where

$$Y_{ik}^* = Q_g^{-1} \circ \hat{F}_k(X_{ik}) = Q_g^{-1}\left(\frac{R_{ik}}{n+1}\right), \qquad i \in \{1, \ldots, n\}, \qquad k \in \{1, \ldots, p\}.$$

The vectors $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_n^*$ are then dependent random variables, from which an empirical process can be constructed. In the light of results by *Ghoudi and Rémillard* [1998, 2004] concerning the asymptotic behavior of such processes, there is every reason to think that when Σ is estimated by $\hat{\Sigma} = (\hat{\sigma}_{k\ell})$, $Z_i = (\mathbf{Y}_i^*)^\top \hat{\Sigma}^{-1} \mathbf{Y}_i^*$ has the same distribution as R^2 for each $i \in \{1, \ldots, n\}$, namely,

$$h(z) = \frac{\pi^{p/2}}{\Gamma(p/2)}\, z^{p/2-1} g(z), \qquad z > 0,$$

with corresponding cumulative distribution function H. This fact is stated here as a conjecture; a formal proof would go well beyond the scope of the present paper.

[56] Assuming that the conjecture is true, a useful procedure for checking that a specific G is a viable choice would then consist of comparing the empirical distribution of the pseudo observations $Z_1,\ldots,Z_n$ to its theoretical counterpart, h. The simplest way to accomplish this visually is to draw a PP plot, whose construction is described next.

[57] A diagnostic PP plot for g: First compute the order statistics $Z_{(1)} < \cdots < Z_{(n)}$ associated with the pseudo observations $Z_1,\ldots,Z_n$. Let also $z_i = (i - 0.5)/n$ for $i \in \{1,\ldots,n\}$. A PP plot for g then consists of the pairs $(z_i, H^{-1}(Z_{(i)}))$ for $i \in \{1,\ldots,n\}$. The plot should be close to a line with unit slope if g is an appropriate choice of metaelliptical copula for the data at hand, and if n is sufficiently large. The z_i are the common Hazen plotting positions. This choice could possibly be refined in subsequent work.

4.3. Goodness-of-Fit Tests

[58] Now suppose that a metaelliptical copula with parameters Σ and g has been selected from a p variate data set. How could one then check the overall quality of the fit with a formal test? Several goodness-of-fit procedures have recently been proposed to this end. They can be divided into three broad classes: (1) tests based on the probability integral transformation of *Rosenblatt* [1952] [e.g., *Breymann et al.*, 2003; *Dobrić and Schmid*, 2007], (2) tests that involve kernel smoothing [e.g., *Fermanian*, 2005; *Panchenko*, 2005; *Scaillet*, 2007], and (3) omnibus tests derived from continuous functionals of the empirical copula process [*Genest et al.*, 2006; *Genest and Rémillard*, 2007]. In this paper, however, the focus is limited to goodness-of-fit tests of the third group. Two reasons motivate this choice:

[59] 1. Tests based on Rosenblatt's transformation involve conditioning on successive components of the random vector and depend on the order in which this conditioning is done.

[60] 2. Although *Scaillet* [2007] has recently streamlined the process, kernel-based goodness-of-fit testing procedures described by *Fermanian* [2005] involve many arbitrary choices (kernel type, window length, weight function, etc.) that make their application cumbersome. Similar criticisms apply to the work of *Panchenko* [2005].

[61] The two tests retained for the present study are described below in separate subsections.

4.3.1. Test Based on the Empirical Copula Process

[62] The first test, originally proposed by *Fermanian* [2005] and implemented by *Genest and Rémillard* [2007], is based on a comparison of the distance between the estimated metaelliptical copula $C_{\hat{\Sigma},g}$ and a rank-based estimate of the underlying copula C requiring no parametric assumption on its form. The latter, according to *Deheuvels* [1979], is defined by

$$C_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\frac{\mathbf{R}_i}{n+1} \le \mathbf{u}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\frac{R_{i1}}{n+1} \le u_1, \ldots, \frac{R_{ip}}{n+1} \le u_p\right), \quad (9)$$

for every $\mathbf{u} = (u_1,\ldots,u_p) \in \mathbb{R}^p$. Specifically, the test statistic considered here is based on the Cramér–von Mises distance

$$S_n = n\int_{(0,1)^p} \left\{C_n(u_1,\ldots,u_p) - C_{\hat{\Sigma},g}(u_1,\ldots,u_p)\right\}^2 dC_n(u_1,\ldots,u_p) = \sum_{i=1}^{n}\left\{C_n\left(\frac{R_{i1}}{n+1},\ldots,\frac{R_{ip}}{n+1}\right) - C_{\hat{\Sigma},g}\left(\frac{R_{i1}}{n+1},\ldots,\frac{R_{ip}}{n+1}\right)\right\}^2.$$

An alternative test based on a Kolmogorov-Smirnov type distance could also be envisaged but is left aside, as it often turns out to be less powerful [see, e.g., *Genest and Rémillard*, 2004; *Genest et al.*, 2006].

[63] An advantage of the goodness-of-fit statistic S_n is that its computation is straightforward, provided that one can rely on an appropriate numerical integration routine for the determination of $C_{\hat{\Sigma},g}(u_1,\ldots,u_p)$ for any choice of $u_1,\ldots,u_p \in (0, 1)$. As the asymptotic distribution of S_n is unwieldy, however, it is preferable to rely on a parametric bootstrap procedure in order to associate a P value to this statistic.

[64] **Algorithm 3**. To compute the P value associated with an observed value of the statistic S_n, fix some large integer N (the larger the better, but $N = 100{,}000$ was taken in the application) and repeat the following steps for every $m \in \{1,\ldots,N\}$:

[65] 1. Generate a random sample $\mathbf{X}^*_{1,m},\ldots,\mathbf{X}^*_{n,m}$ from distribution $\mathcal{E}_p(0, \hat{\Sigma}, g)$ and compute their associated rank vectors $\mathbf{R}^*_{1,m},\ldots,\mathbf{R}^*_{n,m}$.

[66] 2. For every $\mathbf{u} \in \mathbb{R}^p$, let

$$C_{n,m}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\frac{\mathbf{R}^*_{i,m}}{n+1} \le \mathbf{u}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\frac{R^*_{i1,m}}{n+1} \le u_1, \ldots, \frac{R^*_{ip,m}}{n+1} \le u_p\right).$$

[67] 3. Determine the sample value $\tau^*_{k\ell,m}$ of Kendall's tau associated with the pairs $(X^*_{1k,m}, X^*_{1\ell,m}),\ldots,(X^*_{nk,m}, X^*_{n\ell,m})$ and set $\hat{\Sigma}_m = (\hat{\sigma}^*_{k\ell,m})$, where $\hat{\sigma}^*_{k\ell,m} = \sin(\pi\tau^*_{k\ell,m}/2)$.

[68] 4. Compute the value of

$$S^*_{n,m} = \sum_{i=1}^{n}\left\{C_{n,m}\left(\frac{R^*_{i1,m}}{n+1},\ldots,\frac{R^*_{ip,m}}{n+1}\right) - C_{\hat{\Sigma}_m,g}\left(\frac{R^*_{i1,m}}{n+1},\ldots,\frac{R^*_{ip,m}}{n+1}\right)\right\}^2.$$

[69] An approximate P value for the test based on the Cramér–von Mises statistic S_n is then given by

$$\frac{1}{N}\sum_{m=1}^{N} \mathbf{1}\left(S^*_{n,m} > S_n\right).$$

[70] The work of *Genest and Rémillard* [2007] guarantees that this parametric bootstrap procedure provides an adequate approximation of the distribution of S_n under the null hypothesis H_0:$C \in (C_{\Sigma,g})$.

4.3.2. Test Based on Kendall's Process

[71] The second test is adapted from the work of *Genest et al.* [2006]. It is based on a Cramér–von Mises distance

between the distribution of the probability integral transformation $H(\mathbf{X}^*)$ of any p variate vector $\mathbf{X}^*$ with cumulative distribution function H, namely,

$$K(w) = \Pr\left\{H\left(\mathbf{X}^*\right) \le w\right\}, \qquad w \in (0,1),$$

and an empirical estimation K_n thereof. Specifically, a goodness-of-fit test statistic of H_0 that avoids the numerically involved computation of $dK_{\hat{\Sigma},g}(w)/dw$ takes the form

$$T_n = n \int_0^1 \left\{K_n(w) - K_{\hat{\Sigma},g}(w)\right\}^2 dK_n(w),$$

where in general

$$K_{\Sigma,g}(w) = \Pr\left\{C_{\Sigma,g}(\mathbf{X}) \le w\right\}$$

is the distribution induced by $C_{\Sigma,g}$, whose computation must usually rely on a numerical integration routine. As for the empirical analogue K_n, it is given by

$$K_n(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(W_i \le w),$$

where for each $i \in \{1,\ldots,n\}$,

$$W_i = C_n\left(\frac{\mathbf{R}_i}{n+1}\right) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left(\mathbf{R}_j \le \mathbf{R}_i\right) = \frac{1}{n}\sum_{j=1}^{n}\prod_{k=1}^{p} \mathbf{1}\left(R_{jk} \le R_{ik}\right).$$

Accordingly,

$$T_n = \sum_{i=1}^{n}\left\{K_n(W_i) - K_{\hat{\Sigma},g}(W_i)\right\}^2 = \sum_{i=1}^{n}\left\{K_n \circ C_n\left(\frac{R_{i1}}{n+1},\ldots,\frac{R_{ip}}{n+1}\right) - K_{\hat{\Sigma},g} \circ C_n\left(\frac{R_{i1}}{n+1},\ldots,\frac{R_{ip}}{n+1}\right)\right\}^2.$$

The computation of T_n is thus barely any harder than that of S_n, provided that one can rely on an appropriate numerical integration routine for the determination of $K_{\hat{\Sigma},g}(w)$ at arbitrary $w \in (0, 1)$. Here again, the use of a parametric bootstrap is recommended for finding the P value associated with the test statistic.

[72] **Algorithm 4.** To compute the P value associated with an observed value of the statistic T_n, fix some large integer N (again, the larger the better, but $N = 100{,}000$ was taken in the application) and repeat the following steps for every $m \in \{1,\ldots, N\}$:

[73] 1. Generate a random sample $\mathbf{X}^*_{1,m}, \ldots, \mathbf{X}^*_{n,m}$ from distribution $\mathcal{E}_p(0, \hat{\Sigma}, g)$ and compute their associated rank vectors $\mathbf{R}^*_{1,m},\ldots,\mathbf{R}^*_{n,m}$.

[74] 2. For each $i \in \{1,\ldots,n\}$, let

$$W_{i,m} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left(\mathbf{R}^*_{j,m} \le \mathbf{R}^*_{i,m}\right) = \frac{1}{n}\sum_{j=1}^{n}\prod_{k=1}^{p} \mathbf{1}\left(R^*_{jk,m} \le R^*_{ik,m}\right).$$

[75] 3. Compute

$$K_{n,m}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(W_{i,m} \le w\right), \qquad w \in (0,1).$$

[76] 4. Determine the sample value $\tau^*_{k\ell,m}$ of Kendall's tau associated with the pairs $(X^*_{1k,m}, X^*_{1\ell,m}), \ldots, (X^*_{nk,m}, X^*_{n\ell,m})$ and set $\hat{\Sigma}_m = (\hat{\sigma}^*_{k\ell,m})$, where $\hat{\sigma}^*_{k\ell,m} = \sin(\pi\tau^*_{k\ell,m}/2)$.

[77] 5. Compute the value of

$$T^*_{n,m} = \sum_{i=1}^{n}\left\{K_{n,m}\left(W_{i,m}\right) - K_{\hat{\Sigma}_m,g}\left(W_{i,m}\right)\right\}^2.$$

[78] An approximate P value for the test based on the Cramér–von Mises statistic T_n is then given by

$$\frac{1}{N}\sum_{m=1}^{N} \mathbf{1}\left(T^*_{n,m} > T_n\right).$$

[79] Once again, the validity of this parametric bootstrap procedure under the null hypothesis $H_0{:}C \in (C_{\Sigma,g})$ is guaranteed by the work of *Genest and Rémillard* [2007]. These authors also suggest a double parametric bootstrap procedure that may prove a useful substitute to the numerical integration of $C_{\hat{\Sigma},g}$ or $K_{\hat{\Sigma},g}$. This technique is described next.

[80] **Double-bootstrap procedure.** Algorithm 3 requires the evaluation at various points of the cumulative distribution function $C_{\tilde{\Sigma},g}$, either with $\tilde{\Sigma} = \hat{\Sigma}$ or $\hat{\Sigma}_m$. Instead of computing these functions by numerical integration, approximate them by empirical distribution functions of the form

$$\hat{C}_{\tilde{\Sigma},g}(u_1,\ldots,u_p) = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\left(\tilde{X}_{r1} \le u_1,\ldots,\tilde{X}_{rp} \le u_p\right),$$

where $\tilde{\mathbf{X}}_1 = (\tilde{X}_{11},\ldots,\tilde{X}_{1p}),\ldots,\tilde{\mathbf{X}}_R = (\tilde{X}_{R1},\ldots,\tilde{X}_{Rp})$ is a random sample of size R from $C_{\tilde{\Sigma},g}$, for some suitably large integer R, say 100,000.

[81] Similarly, algorithm 4 requires the evaluation at various points of the cumulative distribution function $K_{\tilde{\Sigma},g}$, with $\tilde{\Sigma} = \hat{\Sigma}$ or $\hat{\Sigma}_m$. Instead of proceeding by numerical integration, approximate them by suitable empirical distribution functions. In other words, begin as above by generating a random sample of size R from metaelliptical copula $C_{\tilde{\Sigma},g}$. Then set

$$\tilde{W}_r = \frac{1}{R}\sum_{r'=1}^{R}\prod_{k=1}^{p} \mathbf{1}\left(\tilde{X}_{r'k} \le \tilde{X}_{rk}\right)$$

for each $r \in \{\mathbf{1},\ldots, R\}$ and let

$$\hat{K}_{\tilde{\Sigma},g}(w) = \frac{1}{R}\sum_{r=1}^{R} 1\left(\tilde{W}_r \le w\right), \quad w \in (0,1).$$

5. Analysis of the Romaine Data

[82] It has already been argued in sections 1 and 2 that a trivariate model is required for the analysis of the data on peak, volume and duration of the annual spring flood for the Romaine River. This section shows how metaelliptical copulas and the associated rank-based inference techniques can be used to this end. A univariate analysis of each of

Table 3. Parameter Estimates for Peak, Volume, and Duration of Spring Flood of the Romaine River, Québec, Canada, 1957–2004

Variable	Mixing Weight	First Component of the Mixture	Second Component of the Mixture
Peak	$\pi_1 = 0.23$	$\alpha_1 = 15.86$	$\alpha_2 = 3.85$
		$\beta_1 = 0.01$	$\beta_2 = 0.01$
Volume	$\pi_2 = 0.60$	$\mu_1 = 4478.51$	$\mu_2 = 5719.20$
		$\sigma_1^2 = 823.16$	$\sigma_2^2 = 489.32$
Duration	$\pi_1 = 1$	$\alpha_1 = 0.45$	—
		$\beta_1 = 41.63$	—

the three marginal distributions is described in section 5.1. Copula modeling of the dependence structure between the three variables is then detailed in section 5.2 and finally, simple consequences of the model are drawn in section 5.3.

5.1. Univariate Analysis

[83] Figure 1 shows the evolution of the three main characteristics of the annual spring flood for the Romaine River in the period 1957–2004. Recall that the series comprise only $n = 47$ data points, because no information is available for 1960. The plots for peak and volume suggest a change of hydrologic regime around 1985; flood duration is apparently unaffected by the latter, however. These observations may be confirmed by comparing data before and after 1985 using Wilcoxon's rank-sum test; P values are 0.0355, 0.0355 and 0.0702 for peak, volume and duration, respectively.

[84] Breaking points in the peak and volume series of the Romaine River have been observed on most streams in the Northern Québec/Labrador region east of the Bersimis (or Betsiamites) River [see, e.g., *Perreault et al.*, 2000; *Rousseau and Slivitzky*, 2003]. This phenomenon, which is possibly a reflection of recent climatic changes, cannot be modeled by traditional univariate distributions. Mixtures provide a suitable alternative; they are used in the sequel. Note in passing that their introduction in the model precludes the use of standard multivariate distributions, thereby providing added motivation for the copula approach. Refer to *Favre et al.* [2004] for further discussion on this point.

[85] For the application at hand, peak and volume were respectively modeled using mixtures of gamma and normal distributions, namely,

$$\pi_1\,\Gamma(\alpha_1,\beta_1) + (1-\pi_1)\,\Gamma(\alpha_2,\beta_2) \qquad \text{and}$$
$$\pi_2\,\mathcal{N}\big(\mu_1,\sigma_1^2\big) + (1-\pi_2)\,\mathcal{N}\big(\mu_2,\sigma_2^2\big),$$

where the $\Gamma(\alpha, \beta)$ density is parameterized in such a way that its expectation is α/β. To guide this selection process, the univariate data were analyzed using the software Hyfran [*Chaire en Hydrologie Statistique*, 2002], which uses Akaike and Bayesian information criteria to choose among 16 different distributions.

[86] Because they are particularly well suited for inference in mixture models, Bayesian methods were employed for parameter estimation [see, e.g., *Titterington et al.*, 1985]. More specifically, Gibbs sampling was used to estimate parameters in a Hidden Markov Chain Model formulation where the latent variables correspond to the distributions envisaged. As for duration, it can be represented adequately by a simple gamma distribution, whose parameters were estimated by standard maximum likelihood.

[87] Point estimates for the parameters of the three marginal distributions are summarized in Table 3. As can be seen from Figure 3, the fit of the selected mixtures for peak and volume is adequate (though not ideal). Goodness of fit was confirmed using the classical Kolmogorov-Smirnov test. For peak, volume and duration, the P values were 0.1221, 0.7577 and 0.7657, respectively.

5.2. Dependence Analysis

[88] Statistical analysis usually begins with appropriate graphics, and dependence modeling is no exception. When looking for a copula representation of association, it was mentioned in section 2 that the most telling graphs are rank scatterplots for pairs of variables taken two at a time. One reason is that when these graphs are rescaled through division by $n + 1$, their points correspond to the supports of the various bivariate margins of Deheuvels' empirical copula (9), which is a consistent estimator of the true underlying copula. (Whether for an empirical or a theoretical copula, bivariate margins are found by setting all but two components of the vector $\mathbf{u}$ equal to 1.)

[89] Rank scatterplots (although unscaled) are displayed in Figure 2 for peak, volume and duration of the annual spring flood for the Romaine River data. They suggest that while the pairs (peak,volume) and (volume, duration) are positively associated, the margin-free dependence between peak and duration is negative. As already discussed in section 2, the same conclusions can be derived from the nonparametric, rank-based dependence measures reported in Table 1.

[90] Given the observed asymmetry in the dependence relation between peak, volume and duration, Archimedean

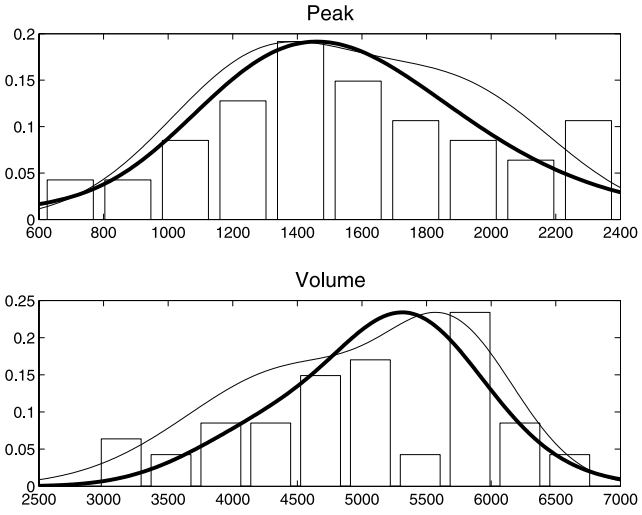

Figure 3. Superposition of histograms and mixtures of normal and gamma densities for peak (m^3/s) and volume (hm^3). The thick lines indicate the gamma mixtures, while the thin lines illustrate the normal mixtures.

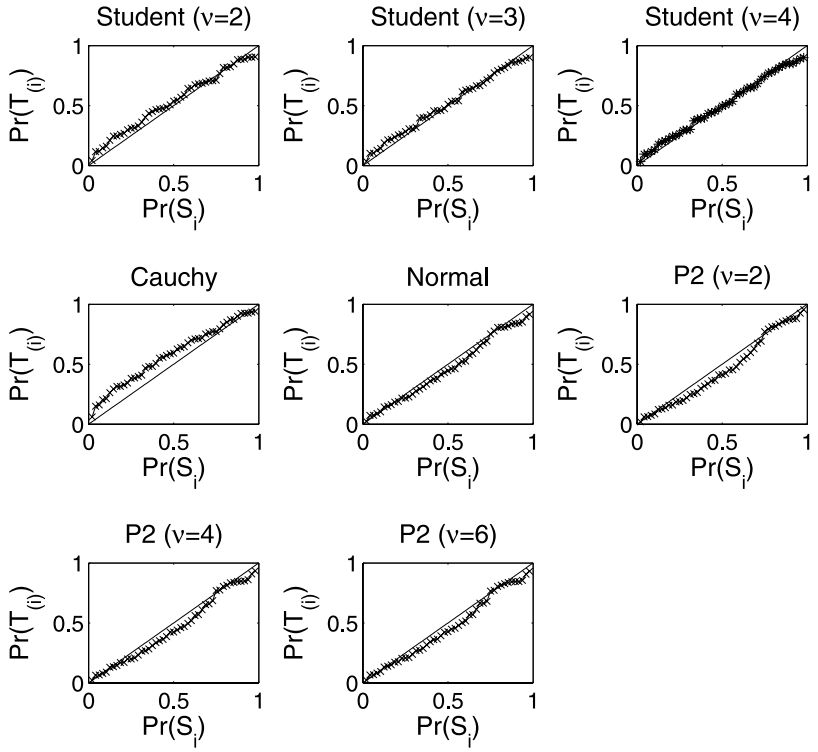

Figure 4. PP plot for the Romaine River data.

copula models are clearly ruled out in this case. Metaelliptical structures of the form (7) may constitute a viable alternative, provided that the parameters g and Σ are suitably chosen.

[91] For the Romaine River data, the procedure described in section 4.1 involves computing $\sin(\pi\tau_{k\ell}/2)$ for $k, \ell \in \{1, 2, 3\}$ with the values of Kendall's tau reported in Table 1 for peak (X_1), volume (X_2) and duration (X_3). This leads to the following estimation of Σ:

$$\hat{\Sigma} = \begin{pmatrix} 1.0000 & 0.5966 & -0.4275 \\ 0.5966 & 1.0000 & 0.2290 \\ -0.4275 & 0.2290 & 1.0000 \end{pmatrix}. \quad (10)$$

This estimate, which is independent of the choice of g, happily turns out to be positive-definite. Its eigenvalues $\lambda_1 = 1.6390$, $\lambda_2 = 1.2142$, $\lambda_3 = 0.1467$ are nonnegative.

[92] As a methodological aside, the following correction proposed by *Rousseeuw and Molenberghs* [1993] could be applied in case $\hat{\Sigma}$ is negative-definite:

[93] 1. Write $\hat{\Sigma}$ in the form $P\Lambda P^{\top}$, where P is orthonormal and $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$ is a diagonal matrix whose entries are the eigenvalues of $\hat{\Sigma}$.

[94] 2. Replace the defective $\hat{\Sigma}$ by $\hat{\Sigma}^* = P\Lambda^* P^{\top}$, where $\Lambda^* = \mathrm{diag}(|\lambda_1|,\ldots,|\lambda_p|)$.

[95] Note that $\hat{\Sigma}^* = \hat{\Sigma}$ whenever the λ_i are positive, so that this procedure has no effect when $\hat{\Sigma}$ is positive-definite to start with.

[96] To choose the generator g, the graphical tool described in section 4.2 may be used. For illustration purposes, eight different choices of g were considered, as per Table 2. They are the trivariate normal, Cauchy and Student with $\nu = 2, 3, 4$ degrees of freedom, as well as the Pearson type II with $\nu = 2, 4, 6$ degrees of freedom.

[97] PP plots for the eight metaelliptical copula models are displayed in Figure 4. They suggest that all models are adequate, except perhaps the Cauchy copula. This observation is vindicated by Table 4, which reports P values for the goodness-of-fit tests S_n and T_n described in section 4.3. Results for T_n are based on 5,000 or 10,000 bootstraps replications, depending on the choice of g; see *Béliveau* [2006] for details.

[98] On the basis of the tests, it is clear that none of the eight copula models considered could be rejected at the 1% level. Only one of them (the Student with $\nu = 4$ degrees of freedom) fails at the 5% level and then again, this occurs only for the test based on T_n.

[99] Unfortunately, such lack of discrimination between dependence structures is typical of small data sets. While distinctions may be introduced as more data are gathered, there is no indication at present, say, that a multivariate normal copula structure (but not a multivariate normal distribution) is inadequate.

[100] As a result, no firm conclusion can be drawn either concerning the existence or not of a tail dependence phenomenon. If the normal copula were selected, for instance, then one would find $\lambda_{12} = \lambda_{13} = \lambda_{23} = 0$ as population values of the upper tail dependence between peak (X_1), volume (X_2) and duration (X_3). Should the Student copula model with ν degrees of freedom were preferred, however, a formula reported by *Frahm et al.*

Table 4. Approximate P Values Obtained by Parametric Bootstrapping for Goodness-of-Fit Testing of Metaelliptical Copula Models of the Romaine River Data

	P Value	
Metaelliptical Copula	S_n (CVM on C_n)	Tn (CVM on Kn)
Normal	0.0892	0.1666
Cauchy	0.1188	0.1050
Student ($\nu = 2$)	0.1264	0.1978
Student ($\nu = 3$)	0.1263	0.1514
Student ($\nu = 4$)	0.1157	0.0450
Pearson ($\nu = 2$)	0.0671	0.1091
Pearson ($\nu = 4$)	0.0795	0.1253
Pearson ($\nu = 6$)	0.0826	0.1261

[2003] implies that in general, the upper tail dependence between variables X_k and X_ℓ is

$$\lambda_{k\ell} = 2 - 2T_{\nu+1}\left(\sqrt{\nu+1}\sqrt{\frac{1-\rho_{k\ell}}{1+\rho_{k\ell}}}\right),\ k,\ell \in \{1,\ldots,p\},$$

where $\mathcal{T}_{\nu+1}$ is the cumulative distribution function of a Student random variable with $\nu + 1$ degrees of freedom.

[101] For the Student copula with $\nu = 2$ degrees of freedom, which provides the best overall fit, this would lead to $\lambda_{12} = 0.448$, $\lambda_{13} = 0.072$, $\lambda_{23} = 0.264$. The least one could say therefore is that caution is advised.

5.3. Consequences of the Dependence Model

[102] In this final paragraph, the importance of taking into account the presence of dependence between variables is illustrated using the Student copula model with $\nu = 2$ degrees of freedom for the Romaine River data.

[103] This model can form the basis for the estimation of various quantities of use for risk analysis, such as conditional probability distributions as well as conditional and joint return periods. These concepts are thoroughly reviewed by *Yue and Rasmussen* [2002] and illustrated by *Salvadori and De Michele* [2004]. It is clear that incorrect use of these notions will lead to a misinterpretation of frequency analysis results.

[104] Consider for example the joint return period

$$T'(x_1,x_2,x_3) = \frac{1}{\Pr(X_1 > x_1, X_2 > x_2, X_3 > x_3)}$$

defined in terms of events involving the peak (X_1), volume (X_2) and duration (X_3) of an annual spring flood. Given that

$$\begin{aligned}\Pr(X_1 > x_1, X_2 > x_2, X_3 > x_3) = 1 &- F_1(x_1) - F_2(x_2) - F_3(x_3)\\ &- F_{123}(x_1,x_2,x_3) + F_{12}(x_1,x_2)\\ &+ F_{13}(x_1,x_3) + F_{23}(x_2,x_3),\end{aligned}$$

it is clear that an estimate of T' (x_1, x_2, x_3) does not depend only on the marginal distributions F_1, F_2 and F_3, but also critically on

$$F_{123}(x_1,x_2,x_3) = C\{F_1(x_1), F_2(x_2), F_3(x_3)\}$$

and on the bivariate margins, namely, $F_{12}(x_1, x_2) = C\{F_1(x_1), F_2(x_2), 1\}$, etc.

[105] In the present case, the margins for peak, volume and duration are as described in Table 3. As for the copula C, a specific choice is more difficult to make, given that none of the eight metaelliptical families could be rejected. As it leads to the largest P value both in terms of S_n and T_n, C is taken here to be the trivariate Student copula with $\nu = 2$ degrees of freedom and $\hat{\Sigma}$ given by (10).

[106] On the basis of this copula model, and using $x_{i[T]}$ to denote the univariate quantile corresponding to X_i for a return period of T years, $i \in \{1, 2, 3\}$, one finds

$$T'\left(x_{1[10\times\gamma]}, x_{2[10\times\gamma]}, x_{3[10\times\gamma]}\right) = 74.3,\ 109.5,\ 136.8 \text{ years}$$

for $\gamma \in \{1, 2, 3\}$. In contrast, if a univariate frequency analysis had been performed under the assumption that the three variables are independent, the joint return period would have been estimated at $T' = \gamma^3 \times 10^3 = 1000$ years, 8000 years or 27,000 years for $\gamma \in \{1, 2, 3\}$, respectively. As one can see, an inappropriate assumption of independence leads in this case to a severe underestimation of the risk associated to this particular event.

[107] As a second illustration, consider a conditional probability of the type $\Pr(X_2 > x_2 \,|X_1 \le x_1, X_3 \le x_3)$. Given x_1, x_2, x_3, this quantity can be expressed as

$$\begin{aligned}\Pr(X_2 > x_2 | X_1 \le x_1, X_3 \le x_3) &= \frac{\Pr(X_2 > x_2, X_1 \le x_1, X_3 \le x_3)}{\Pr(X_1 \le x_1, X_3 \le x_3)}\\ &= \frac{F_{13}(x_1,x_3) - F_{123}(x_1,x_2,x_3)}{F_{13}(x_1,x_3)}\end{aligned}$$

with the same notations as above. While

$$\Pr\left(X_2 > x_{2[10\times\gamma]} | X_1 \le x_{1[10\times\gamma]}, X_3 \le x_{3[10\times\gamma]}\right) = \frac{1}{10\gamma},\ \gamma \in \{1,2,3\}$$

under the assumption of mutual independence, the copula model leads to much more realistic estimates, namely .02, .013 and .010 for $\gamma \in \{1, 2, 3\}$, respectively. This comes to show, once again, the detrimental effects of taking for granted stochastic independence in situations where this assumption is clearly inappropriate.

6. Conclusion

[108] This paper has shown how metaelliptical copulas can help in modeling the dependence structure of random vectors when observed differences between their bivariate margins preclude the use of exchangeable copula families, e.g., the Archimedean class. Peak, volume and duration of the annual spring flood for the Romaine River were used to illustrate rank-based estimation and goodness-of-fit techniques for this broad extension of the multivariate normal distribution.

[109] The analysis of the data at hand suggests that in view of the short length of the series, any of the eight metaelliptical copula models considered here could be used for prediction purposes. Only with additional evidence could one hope to distinguish between these dependence structures. At present, the P values reported in Table 4 suggest that the Student copula with $\nu = 2$ degrees of

freedom and $\hat{\Sigma}$ given by (10) would be the most sensible choice for computing, e.g., joint return periods or conditional probabilities of events of interest as computed in section 4.3. Here as in many other statistical contexts, however, the importance of performing a sensitivity analysis (and in particular, not taking independence for granted) cannot be overly emphasized.

[110] To the recent convert, the fact that the multivariate normal copula is a realistic model for the dependence of the Romaine River data may come as somewhat of a disappointment. This may lead him/her to think that in this specific application at least, the copula approach has not provided a significant improvement over standard hydrologic modeling using the multivariate normal distribution. On second thought, however, he/she will come to realize that the combination of a normal copula with heterogeneous margins, including mixture distributions!, is a far cry from the traditional Gaussian model.

[111] **Acknowledgments.** Partial funding in support of this work was provided by Hydro-Québec, the Natural Sciences and Engineering Research Council of Canada, the Fonds Québécois de la Recherche sur la Nature et les Technologies, and the Institut de Finance Mathématique de Montréal.

References

Abdous, B., C. Genest, and B. Rémillard (2005), Dependence properties of meta-elliptical distributions, in *Statistical Modeling and Analysis for Complex Data Problems*, edited by P. Duchesne and B. Rémillard, pp. 1–15, Springer, New York.

Béliveau, J. (2006), Analyse fréquentielle multivariée de la pointe, du volume et de la durée de crue, M. S. thesis, Univ. Laval, Quebec, Que., Canada.

Breymann, W., A. Dias, and P. Embrechts (2003), Dependence structures for multivariate high-frequency data in finance, *Quant. Finan.*, *3*, 1–14.

Chaire en Hydrologie Statistique (2002), Hyfran, logiciel pour l'analyse fréquentielle en hydrologie, technical report, INRS-ETE, Univ. of Quebec, Que., Canada.

Cherubini, U., E. Luciano, and W. Vecchiato (2004), *Copula Methods in Finance*, John Wiley, Hoboken, N. J.

Deheuvels, P. (1979), La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance, *Acad. R. Bel. Bull. Class. Sci.*, *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.

De Michele, C., G. Salvadori, M. Canossi, A. Petaccia, and R. Rosso (2005), Bivariate statistical approach to check adequacy of dam spillway, *J. Hydrol. Eng.*, *10*, 50–57.

Dobrić, J., and F. Schmid (2007), A goodness of fit test for copulas based on Rosenblatt's transformation, *Comput. Stat. Data Anal.*, *51*, 4633–4642.

Fang, H.-B., K.-T. Fang, and S. Kotz (2002), The meta-elliptical distributions with given marginals, *J. Multivariate Anal.*, *82*, 1–16, (Corrigendum, *J. Multivariate Anal., 94*, 222–223, 2005.).

Fang, K. T., S. Kotz, and K. W. Ng (1990), *Symmetric Multivariate and Related Distributions*, CRC Press, Boca Raton, Fla.

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.

Fermanian, J.-D. (2005), Goodness-of-fit tests for copulas, *J. Multivariate Anal.*, *95*, 119–152.

Frahm, G., M. Junker, and R. Schmidt (2003), Elliptical copulas: Applicability and limitations, *Stat. Probab. Lett.*, *63*, 275–286.

Genest, C., and A.-C. Favre (2007), Everything you always wanted to know about copula modeling but were afraid to ask, *J. Hydrol. Eng.*, *12*, 347–368.

Genest, C., and B. Rémillard (2004), Tests of independence and randomness based on the empirical copula process, *Test*, *13*, 335–369.

Genest, C., and B. Rémillard (2007), Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models, *Ann. Inst. Henri Poincare*, in press.

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

Genest, C., and F. Verret (2005), Locally most powerful rank tests of independence for copula models, *J. Nonparametric Stat.*, *17*, 521–539.

Genest, C., and B. J. M. Werker (2002), Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models, in *Distributions With Given Marginals and Statistical Modelling*, edited by C. M. Cuadras, J. Fortiana, and J. A. Rodríguez Lallena, pp. 103–112, Springer, New York.

Genest, C., K. Ghoudi, and L.-P. Rivest (1995), A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika*, *82*, 543–552.

Genest, C., J.-F. Quessy, and B. Rémillard (2006), Goodness-of-fit procedures for copula models based on the probability integral transformation, *Scand. J. Stat.*, *33*, 337–366.

Ghoudi, K., and B. Rémillard (1998), Empirical processes based on pseudo-observations, in *Asymptotic Methods in Probability and Statistics*, edited by B. Szyskowicz, pp. 171–197, Elsevier, New York.

Ghoudi, K., and B. Rémillard (2004), Empirical processes based on pseudo-observations. II. The multivariate case, *Fields Inst. Commun.*, *44*, 381–406.

Grimaldi, S., and F. Serinaldi (2006), Asymmetric copula in multivariate flood frequency analysis, *Adv. Water Resour.*, *29*, 1155–1167.

Hult, H., and F. Lindskog (2002), Multivariate extremes, aggregation and dependence in elliptical distributions, *Adv. Appl. Probab.*, *34*, 587–608.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, CRC Press, Boca Raton, Fla.

Kendall, M. G. (1955), *Rank Correlation Methods*, Griffin, London.

Lee, A. J. (1990), *U-Statistics: Theory and Practice*, CRC Press, Boca Raton, Fla.

Lindskog, F., A. J. McNeil, and U. Schmock (2003), A note on Kendall's tau for elliptical distributions, in *Credit Risk: Measurement, Evaluation and Management*, edited by G. Bol et al., pp. 149–156, Physica, Heidelberg, Germany.

Marsaglia, G. (1972), Choosing a point from the surface of a sphere, *Ann. Math. Stat.*, *43*, 645–646.

Marshall, A. W., and I. Olkin (1988), Families of multivariate distributions, *J. Am. Stat. Assoc.*, *83*, 834–841.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer, New York.

Panchenko, V. (2005), Goodness-of-fit test for copulas, *Physica A*, *355*, 176–182.

Perreault, L., E. Parent, J. Bernier, B. Bobée, and M. Slivitzky (2000), Retrospective multivariate Bayesian change-point analysis: A simultaneous single change in the mean of several hydrological sequences, *Stochastic Environ. Res. Risk Assess.*, *14*, 243–261.

Poulin, A., D. Huard, A.-C. Favre, and S. Pugin (2007), On the importance of the tail dependence in bivariate frequency, *J. Hydrol. Eng.*, *12*, 394–403.

Rao, C. R. (1981), *Linear Statistical Inference*, John Wiley, Hoboken, N. J.

Rosenblatt, M. (1952), Remarks on a multivariate transformation, *Ann. Math. Stat.*, *23*, 470–472.

Rousseau, A., and M. Slivitzky (2003), Relationships between annual runoff variability and the Arctic Oscillation Index in the Northern Québec/Labrador peninsula, in *Proceedings of the 18th Stanstead Seminar: Climate Variability and Predictability From Seasons to Decades*, edited by J. Derome et al., pp. 118–121, Can. CLIVAR Res. Network, Montreal, Que., Canada.

Rousseeuw, P. J., and G. Molenberghs (1993), Transformation of nonpositive semidefinite correlation matrices, *Commun. Stat. Theory Methods*, *22*, 965–984.

Salvadori, G., and C. De Michele (2004), Frequency analysis via copulas: Theoretical aspects and applications to hydrological events, *Water Resour. Res.*, *40*, W12511, doi:10.1029/2004WR003133.

Salvadori, G., and C. De Michele (2006), Statistical characterization of temporal structure of storms, *Adv. Water Resour.*, *29*, 827–842.

Scaillet, O. (2007), Kernel based goodness-of-fit tests for copulas with fixed smoothing parameters, *J. Multivariate Anal.*, *98*, 533–543.

Schmidt, R. (2002), Tail dependence for elliptically contoured distributions, *Math. Method. Oper. Res.*, *55*, 301–327.

Shih, J. H., and T. A. Louis (1995), Inferences on the association parameter in copula models for bivariate survival data, *Biometrics*, *51*, 1384–1399.

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

Yue, S., and P. Rasmussen (2002), Bivariate frequency analysis: Discussion of some useful concepts in hydrological application, *Hydrol. Processes*, *16*, 2881–2898.

Zhang, L., and V. P. Singh (2006), Bivariate flood frequency analysis using the copula method, *J. Hydrol. Eng.*, *11*, 150–164.

J. Béliveau, Direction des Politiques et Programmes, Ministère de l'Éducation, du Loisir et du Sport, 1035, rue de la Chevrotière, Québec, QC, Canada G1R 5A5. (julie.beliveau@mels.gouv.qc.ca)

A.-C. Favre, Chaire en hydrologie statistique, INRS-ETE, 490, rue de la Couronne, Québec, QC, Canada G1K 9A9. (anne-catherine_favre@ete.inrs.ca)

C. Genest, Département de mathématiques et de statistique, Université Laval, Québec, QC, Canada G1K 7P4. (christian.genest@mat.ulaval.ca)

C. Jacques, Service du devéloppement et des infrastructures stratégiques, Ministère des Affaires Municipales et des Régions, 10, rue Pierre-Olivier-Chauveau, Québec, QC, Canada G1R 4J3. (christiane.jacques@mamr.gouv.qc.ca)