Use of a Gaussian copula for multivariate extreme value analysis: some case studies in hydrology

Benjamin Renard, M. Lang

Author-produced version of the article published in Advances in Water Resources, 2007, 30, 897-912.
The original publication is available at http://www.sciencedirect.com/
doi : 10.1016/j.advwatres.2006.08.001

USE OF A GAUSSIAN COPULA FOR MULTIVARIATE EXTREME VALUE ANALYSIS: SOME CASE STUDIES IN HYDROLOGY.

B. RENARD[1], M. LANG[1]

(1) *Cemagref Centre de Lyon, U.R. Hydrologie-Hydraulique, 3 bis Quai Chauveau, CP 220, 69336 Lyon cedex 09, France.*

telephone:33 4 72 20 87 72
fax:33 4 78 47 78 75

e-mail : renard@lyon.cemagref.fr, lang@lyon.cemagref.fr

Abstract: Risk assessment requires a description of the probabilistic properties of hydrological variables. In a number of cases, this description is made on a single variable, whereas most hydrological events are intrinsically multivariate. In this context, copulas have recently received attention in order to derive a multivariate frequency analysis. After a reminder of the general results in the field of multivariate extreme value theory, the paper gives a description of a very simple copula, the Gaussian copula. Four case studies demonstrate its usefulness in the contexts of field significance determination, regional risk analysis, Discharge-Duration-Frequency (QdF) models with design hydrograph derivation and regional frequency analysis. The limitations and potential errors related to this statistical tool are also highlighted.

Keywords: multivariate analysis; extreme value analysis; extremes dependence; copula; field significance; regional frequency analysis; QdF models; design hydrographs; asymptotic properties; risk assessment.

I. Introduction

Extreme value theory (EVT) is widely used by hydrologists, especially for flood and drought mitigation. The severity of a hydrological extreme event is expressed as a non-exceedance probability, or equivalently, in terms of return period. This can be used to locate an observed event at a probabilistic scale and the theory can be used for extrapolation, i.e. for the determination of the probability of observing an event, even outside the range of observations. In some cases, hydrologists can be interested in several hydrological variables, which would lead to use of a multivariate statistical tool. Examples of such a situation in the flood mitigation field include the following:

- A flood event can be studied in different ways. The most commonly assessed trait is the peak flow of the event, but the volume or threshold exceedance duration can also be of interest. As an example, an event with a peak of a hundred-years return period could be less damaging than an event with a ten-years return period both in peak and volume. Unfortunately, these traits are not independent, thus preventing each variable from being studied separately. The multivariate distribution of the triplet (peak, volume, duration) is needed. Such an analysis has been undertaken by Adamson et al. [1] or Grimaldi and Serinaldi [23] as an example.
- A flood event at a river confluence can be the result of high discharges in only one or both of the upstream flows. If the discharges of these two streams are independent, the probabilistic behavior of the downstream flow, equal to the sum of the two upstream ones, can be obtained by convolution. Nevertheless, both tributaries can tend to have simultaneous high flows. In this case, the preceding calculation can lead to a strong underestimation of the risk. Once again, the multivariate distribution of the two upstream flows is needed. Some examples are provided by Mousavi [46], Le Clerc and Lang [35] and Favre et al. [17]. Coles [6] describes the underlying methodology.
- Water resources managers usually have to deal with several rivers in a given geographical area. A currently observed phenomenon is that a ten-year flood event is observed almost every year at a regional scale: the at-site non-exceedance probability is not suitable at the regional scale [62]. This phenomenon can easily be quantified in the case of M independent sites: the probability of observing at least one event with non-exceedance probability p is equal to $1-p^M \geq 1-p$. In contrast, if all sites are perfectly correlated, then the regional and the local probabilities are equal. Between these two extreme cases, which are never encountered in practice, the regional behavior of flood events is related to the dependence between stations, which has to be taken into account through a multivariate distribution.

Multivariate extreme value theory (MEVT) is studied by mathematicians (see Coles [6] for a review of the topic). Unfortunately, theoretical results are often difficult to use in practice, compared to the classical EVT. Additional hypotheses therefore have to be made to take into account dependence between variables. In this context, copulas have recently received particular attention [4, 13, 17, 21, 23, 54, 57], because they are relatively easy to handle and they can be used in a wide range of situations. In counterpart, they are based on a model of the dependence structure, which has to be checked for adequacy.

The aim of this article is to present some examples involving a particular copula, the Gaussian copula. In the first part, a review of some results of MEVT will be given, with a general description of copulas, and a more detailed description of the Gaussian copula (section II). The first application deals with the problem of the field significance determination in multi-

testing problems (section III.1). The second example deals with regional risk estimation, as presented in the introduction (section III.2). In section III.3, the Discharge-Duration-Frequency (QdF) methodology will be studied, and a copula will be used to compute the return period of a design hydrograph. The problem of intersite dependence and its implications for regional frequency analysis will be explored in section III.4. Examples where a Gaussian copula is not able to model properly the dependence structure of observations will also be presented (section III.5). Finally, the results obtained in this paper will be discussed (section IV), before giving some conclusions about the advantages and drawbacks in using a Gaussian copula, and proposing some perspectives for improving multivariate hydrological extreme events analysis (section V).

II. Multivariate extreme value theory

II.1. Some results of MEVT for componentwise maxima

Let $(X_1,...,X_n)$ be independent and identically distributed variables. Let M_n denote the maximum of these variables, $M_n = \max_{i=1,...,n}(X_i)$. Under regularity conditions, it can be proven [18] that the only possible limit distribution of M_n, normalized by suitable values, is the general extreme value distribution (GEV), whose density is:

$$f(x;\alpha,\beta,\xi) = \frac{1}{\alpha}\left(1-\frac{\xi(x-\beta)}{\alpha}\right)^{\frac{1}{\xi}-1}\exp\left(-\left(1-\frac{\xi(x-\beta)}{\alpha}\right)^{\frac{1}{\xi}}\right) \quad (1)$$

This theorem justifies the use of the GEV distribution to model annual maximum discharges, for example. In the multivariate case, such a theorem is also available. For clarity the bivariate case will be described. Let $(X_i,Y_i)_{i=1,...,n}$ be independent and identically distributed vectors. Let M_n denote the 2-dimensional vector of the componentwise maxima, $M_n = (M_n^X, M_n^Y) = \left(\max_{i=1,...,n}(X_i), \max_{i=1,...,n}(Y_i)\right)$. As previously, the only possible limit distribution of M_n has the form [6]:

$$\begin{aligned} G(x,y) &= \Pr(M_n^x/n \le x, M_n^y/n \le y) \\ &= \exp\{-V(\tilde{x},\tilde{y})\} \end{aligned}$$

where:

$$\tilde{x} = \left(1-\frac{\xi_x}{\alpha_x}(x-\beta_x)\right)^{-1/\xi_x}, \tilde{y} = \left(1-\frac{\xi_y}{\alpha_y}(y-\beta_y)\right)^{-1/\xi_y} \quad (2)$$

$$V(x,y) = 2\int_0^1 \max\left(\frac{w}{x},\frac{1-w}{y}\right)dH(w)$$

and H is a distribution on [0;1] with a mean of ½.

Unfortunately, it appears that the limit distribution cannot be expressed in a classical parameterized way: an infinity of H distributions is able to validate the mean constraint. The choice of this function is the main difficulty of multivariate extreme value analysis. In practice, $G(x,y)$ can be chosen in a parametric family of distributions, which will be able to model a wide range of dependence structures, ranging from independence to total dependence. An example of such a family is the logistic family [6]. Up to this point, it is important to notice that this choice is a hypothesis which is not based on theoretical results, in

contrast to the use of the GEV in the univariate case. Consequently, it is necessary to check the adequacy of the dependence model and to keep in mind that the overall uncertainty is made up of the uncertainty related to the estimation of the parameters, which can be quantified by standard statistical methods, and of the modeling uncertainty, which is more difficult to assess.

II.2. Copulas

Copulas are alternative tools for dealing with multivariate extremes, and have become very popular in recent years. Let $(X^{(1)},\ldots,X^{(d)})$ be a d-dimensional random vector with a probability distribution of:

$$F(x_1,...,x_d) = \Pr(\{X^{(1)} \le x_1\} \cap ... \cap \{X^{(d)} \le x_d\}) \tag{3}$$

A copula is a function c verifying:

$$\begin{aligned} &c:[0;1]^d \mapsto [0;1] \\ &\text{and} \\ &F(x_1,...,x_d) = c(F_1(x_1),...,F_d(x_d)) \end{aligned} \tag{4}$$

The c function is used to model dependence between variables, while marginal distributions can be described in an usual way, with for example GEV distributions. Consequently, the copula function can be used in order to derive a multivariate distribution with the desired marginal distributions. Conversely, let F be a multivariate distribution with marginal distributions $F_1,\ldots,F_d$. It can then be proven that c exists [56], which means that any multivariate distribution can be written in the form of equation (4). Moreover, if the marginal distributions F_i are continuous, then c is unique.

Nevertheless, as with the H distribution, no theoretical results are available to determine the copula: given marginal distributions, an infinity of multivariate joint distributions can be derived. Once again, a parametric family has to be chosen to model the dependence (see Favre et al. [17] for a review). As an illustration, Archimedean copulas can be constructed as follows. Let $\varphi:[0;1]\to[0;+\infty]$ be a continuous decreasing function such that $\varphi(1)=0$ and $(-1)^k d^k \varphi^{-1}(t)/dt^k \ge 0$ for all t in $]0;+\infty[$ and k=1,…,d, where d is the number of dimensions. The following function is then a copula:

$$\begin{aligned} &c:[0;1]^d \mapsto [0;1] \\ &c(x_1,...,x_d) = \begin{cases} \varphi^{-1}\left(\sum_{i=1}^{d} \varphi(F_i(x_i))\right) & \text{if } \sum_{i=1}^{d} \varphi(F_i(x_i)) \le \varphi(0) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

Classical choices for the φ generator function include $\varphi(t) = t^{-\alpha} - 1$ (Clayton copula), $\varphi(t) = \log\left(\frac{e^{\alpha t}-1}{e^{\alpha}-1}\right)$ (Frank copula) or $\varphi(t) = (-\log t)^{\alpha}$ (Gumbel-Hougaard copula). The α parameter thus summarizes the dependence. Although such copulas are very useful in the bivariate case, they are more problematic with a high number of dimensions, because a single parameter is not sufficient to describe a random vector with contrasted levels of dependence between marginal components. It is thus necessary to generalize the previous generator functions, by including additional parameters. Grimaldi and Serinaldi [23] provide an example of such a generalization. An alternative family may be easier to handle in highly dimensional cases and comprises the Elliptical copulas. These copulas use a symmetric and positive definite matrix in order to model dependence. The elements of this matrix can be interpreted as dependence measures between couples of variables, leading to an analogy with

the correlations used in the case of a multivariate Gaussian distributions. The most well known elliptical copulas are the Gaussian copula, which will be described in details later, and the Student's copula.
Whatever the family chosen, the number of parameters quickly grows with dimension. Estimation using standard methods such as maximum likelihood can be impossible because of numerical difficulties. A common estimation scheme consists in estimating marginal parameters separately as a first step, then estimating copula parameters knowing marginal estimations. On the other hand, the efficiency of estimators is not ensured.
Although appealing, the copula theory also suffers from a number of drawbacks, essentially because of its lack of connection with standard multivariate extreme value theory. Although specific copulas have been proposed to account for the theoretical properties of extremes [4, 22], additional developments are still needed, especially in the field of spatial extremes. This problem may be a limit for extrapolation, i.e. for the computation of very low probabilities, corresponding to multivariate events lying outside the observations range. Mikosh [45] thus lists a number of questions that remain unsolved, and provides a skeptical point of view about the use of copulas.

II.3. The Gaussian copula

The Gaussian copula is a member of the Elliptical copulas family. The dependence is modeled by means of a symmetric and definite positive matrix, whose elements are used to describe the dependence between couples of variables. This model is thus convenient when the number of dimensions is more than two or three, and has been widely studied in finance (see Cherubini et al. [5] and references therein). In the hydrology field, the meta-Gaussian density studied by Kelly and Krzysztofowicz [32] and Herr and Krzysztofowicz [25] can also be viewed as a Gaussian copula.
The Gaussian copula is defined as follows:

$$c(u_1,...,u_d) = \Phi_d\left(\phi^{-1}(u_1),...,\phi^{-1}(u_d)\right) \tag{6}$$

where ϕ is the cdf of the standard normal distribution *N(0,1)* and Φ_d is the cdf of a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}$.
In other words, the multivariate cumulative distribution of the data will have the following form:

$$\begin{aligned} F(x_1,...,x_d) &= c(F_1(x_1),...,F_d(x_d)) \\ &= \Phi_d\left(\phi^{-1}(F_1(x_1)),...,\phi^{-1}(F_d(x_d))\right) \end{aligned} \tag{7}$$

or equivalently, the multivariate density can be written as:

$$f(x_1,...,x_d) = f_1(x_1)\times...\times f_d(x_d)\times$$
$$|\Sigma|^{-1/2}\exp\left\{-\frac{\left[\phi^{-1}\left(F_1(x_1)\right),...,\phi^{-1}\left(F_d(x_d)\right)\right]\left[\Sigma^{-1}-I\right]\left[\phi^{-1}\left(F_1(x_1)\right),...,\phi^{-1}\left(F_d(x_d)\right)\right]^T}{2}\right\} \tag{8}$$

The main advantage of this copula is its simplicity: once the data have been transformed by $\phi^{-1}(F_i(.))$, the well known multivariate normal distribution is used to calculate probabilities.
Figure 1 illustrates the principle of the copula, while Figure 2 shows the shape of the multivariate distribution obtained in a two-dimensional case. It also shows the limit of the Gaussian copula: with GEV marginal distributions, the distribution of the data has to be suitably described by this V-shaped dependence. In particular, a Gaussian copula will not be able to model properly more complex dependence structures. As an example, dependence between extreme variables may depend on marginal values. In other cases, it may be inappropriate to model the dependence of transformed variables by simple correlations. It is

therefore necessary to check the adequacy of the dependence structure implied by the Gaussian copula. Standard model diagnosis used in multivariate Gaussian analysis can be used for this purpose. As an example, Chi-square plots provide a graphical diagnosis for multivariate normality. The principle can be described as follows: let $(\tilde{X}^{(1)},...,\tilde{X}^{(d)})$ denote the *d*-dimensional variable of interest, transformed toward standardized marginal normality thanks to $\phi^{-1}(F_i(.))$. Under the assumption of multivariate normality $N(\boldsymbol{0};\boldsymbol{\Sigma})$, the following quadratic form should fit a χ^2_d distribution:

$$D = (\tilde{X}^{(1)},...,\tilde{X}^{(d)})\boldsymbol{\Sigma}^{-1}(\tilde{X}^{(1)},...,\tilde{X}^{(d)})^T \quad (9)$$

In practice, the data are transformed by a normal-score transformation, in order to avoid departure from the expected distribution which may be due to marginal estimation errors, and the correlation matrix is estimated empirically on the transformed data. The transformed data are then used to construct a sample of d_i, by:

$$d_i = (\tilde{x}_i^{(1)},...,\tilde{x}_i^{(d)})\hat{\boldsymbol{\Sigma}}^{-1}(\tilde{x}_i^{(1)},...,\tilde{x}_i^{(d)})^T \quad (10)$$

A QQ-plot can finally be constructed: if $d_{(i)}$ denotes the i^{th} sorted value of $(d_i)_{i=1,...,n}$ and $\chi^2_d(\alpha)$ the α-quantile of a Chi-square distribution with *d* degrees of freedom, then the points $\left(d_{(i)};\chi^2_d\left(\frac{i-0.5}{n}\right)\right)$ should remain close to the *y=x* line. Alternative diagnostic tools may be used, for example by using univariate goodness-of-fit tests based on linear combinations of the marginal values (see *e.g.* Mardia [40] for additional methods).

The next step consists in estimating parameters. The estimation scheme presented in the previous section can be used to estimate first the marginal parameters, then the $\boldsymbol{\Sigma}$ covariance matrix. Phoon et al. [49] proposed an alternative method, which allows the marginal and dependence parameters to be estimated independently:

1. Estimation of Spearman's rank correlations $r_{i,j}$ between pairs of variables. These correlations are invariant by monotonic transformation.
2. Transformation toward Pearson correlation coefficients by $\rho_{i,j} = 2\sin(\frac{\pi}{6}r_{i,j})$
3. Estimation of marginal parameters (Maximum likelihood for instance).

The drawback of this method is that the obtained $\boldsymbol{\Sigma}$ covariance matrix is not ensured to be non-negative definite. If this is not the case, $\boldsymbol{\Sigma}$ can be estimated by the classical observed covariance matrix between transformed data $\tilde{X}$ by $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\tilde{X}^T\tilde{X}$.

To conclude with this presentation, notice that a Gaussian copula is an easy tool to simulate data with prescribed marginal distributions and correlations: as a first step, a multivariate normal data set is simulated from $\Phi_d(\boldsymbol{0},\boldsymbol{\Sigma})$, where the elements of the diagonal of $\boldsymbol{\Sigma}$ are equal to one. Marginal samples are then transformed by $F_i^{-1}(\phi(.))$.

III. Use of the Gaussian copula in Hydrology

III.1. Field significance

Field significance determination is a problem encountered in multi-testing studies. It has been studied by Livezey and Chen [37], Lettenmaier et al. [36] , Douglas et al. [15], and Yue and Wang [63]. In the hydro-meteorological field, a major preoccupation is the determination of

the impacts of climatic change on various variables. The method normally consists in testing a number of stations for stationarity. As an example, suppose that 100 stations are tested with a risk equal to 0.05. Under the hypothesis that all stations are stationary, approximately five stations should be detected as non-stationary, because of the 5% risk. But what would be the conclusion with 6 significant results? In other words, what is the minimum number of locally significant results to conclude, with a risk α', that all these results cannot be due to chance? In order to answer this question, the distribution of N, the number of locally significant tests under the hypothesis that all series are stationary, is needed. This distribution can be calculated theoretically as follows: let $\boldsymbol{X}^{(i)}$ denote the data recorded at site i, i=1,…, p. Assume that all series are tested by comparing a test statistic $S_i=g(\boldsymbol{X}^{(i)})$ with a critical value c. Let Ω_k denote the set of all possible combinations of k elements among p. The distribution of N can then be derived as:

$$\begin{aligned}
\Pr(N=k) &= \sum_{\Omega_k} \Pr\left(S_{n_1} > c,...,S_{n_k} > c, S_{n_{k+1}} \le c,...,S_{n_p} \le c\right) \\
&= \sum_{\Omega_k} \Pr\left(g(\boldsymbol{X}^{(n_1)}) > c,...,g(\boldsymbol{X}^{(n_k)}) > c, g(\boldsymbol{X}^{(n_{k+1})}) \le c,...,g(\boldsymbol{X}^{(n_p)}) \le c\right) \\
&= \sum_{\Omega_k} \Pr\left(\boldsymbol{X}^{(n_1)} \in g^{-1}(]c;+\infty[),...,\boldsymbol{X}^{(n_k)} \in g^{-1}(]c;+\infty[), \boldsymbol{X}^{(n_{k+1})} \in g^{-1}(]-\infty;c]),...,\boldsymbol{X}^{(n_p)} \in g^{-1}(]-\infty;c])\right) \\
&= \sum_{\Omega_k} \Pr\left(\boldsymbol{X} \in \Re_{\Omega_k}\right)
\end{aligned} \tag{11}$$

where $\Re_{\Omega_k}$ is defined by:

$$\left\{\boldsymbol{X}=\left(\boldsymbol{X}^{(1)},...,\boldsymbol{X}^{(p)}\right) \middle| \boldsymbol{X}^{(n_1)} \in g^{-1}(]c;+\infty[),...,\boldsymbol{X}^{(n_k)} \in g^{-1}(]c;+\infty[), \boldsymbol{X}^{(n_{k+1})} \in g^{-1}(]-\infty;c]),...,\boldsymbol{X}^{(n_p)} \in g^{-1}(]-\infty;c])\right\} \tag{12}$$

This equation shows that the multivariate distribution of the tested series is needed, except in the case of independence. In this case:

$$\begin{aligned}
\Pr(N=k) &= \sum_{\Omega_k} \Pr\left(S_{n_1} > c,...,S_{n_k} > c, S_{n_{k+1}} \le c,...,S_{n_p} \le c\right) \\
&= \sum_{\Omega_k} \Pr\left(S_{n_1} > c\right)...\Pr\left(S_{n_k} > c\right)\Pr\left(S_{n_{k+1}} \le c\right)...\Pr\left(S_{n_p} \le c\right) \\
&= \sum_{\Omega_k} \alpha^k (1-\alpha)^{p-k} \\
&= C_p^k \alpha^k (1-\alpha)^{p-k}
\end{aligned} \tag{13}$$

If all sites are independent, N follows a binomial distribution. With the preceding example, the minimum number of locally significant results to ensure a 5% field significance is thus equal to nine under the independence hypothesis, because $p(N \le 8) = 0.937$ and $p(N \le 9) = 0.972$.

If the tested series are dependent, a Gaussian copula can be used to take into account spatial correlations in field significance determination, in the following way:

- parameters estimation
- do i=1,…, M
    - simulation of a new multivariate data set thanks to the estimated copula
    - test of each series for stationarity
    - computation of N_i, the number of significant results

The simulated data series are all stationary, and agree with the observed correlations and marginal distributions. The sample of (N_i) can thus be used to approximate the distribution of N, and to compute a critical value.

This method was applied to a set of 13 hydrometric stations in the North-East of France. Between 39 and 84 years of daily discharges are available. Annual maxima were extracted and a Mann-Kendall test [33, 39] was applied to the 13 series obtained. Five stations showed a significant trend at risk 5% (Figure 3). The preceding procedure was then applied in order to evaluate the field significance of these five significant results. Marginal distributions are assumed to be Gumbel distributions, whose parameters are estimated with the method of moments on all available years. The $\boldsymbol{\Sigma}$ correlation matrix was directly estimated on transformed data by $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, because the Phoon method based on rank correlations led to a non-definite positive matrix. Only years shared by all stations were taken into account for this estimation (30 years). Estimated correlations ranged from 0.051 to 0.976. The adequacy of the Gaussian copula model was evaluated with the QQ-plot described in section II.3 (Figure 4), which shows no strong departure from the expected distribution. The distribution of N was finally approximated with M=1000 simulations, as shown by Figure 5. The decision depends on the combination of a local risk and a regional risk, which are not necessarily equal. Different combinations are summarized in Table 1. As an example, the five significant changes obtained with a 5% test were also regionally significant at a risk of 5%, but were not at risk 1%.

III.2. Regional risk estimation

In this section, the aim will be to estimate the probability of observing at least one event of given return period T, in a set of d stations, as explained in the introduction. The return period corresponding to this probability will be called the regional return period T_R. A similar analysis can be found in the paper by Troutman and Karlinger [62]. Let $X^{(i)}$ denote the annual maximum discharge of station i. If a Gaussian copula is suitable for modeling the multivariate data set, then this probability can be computed as follows:

$$\begin{aligned}
P(\text{at least one event with return period T}) &= 1 - P(\text{no event with return period T}) \\
&= 1 - P(\{X^{(1)} \le q_T^{(1)}\} \cap \ldots \cap \{X^{(d)} \le q_T^{(d)}\}) \\
&= 1 - F(q_T^{(1)}, \ldots, q_T^{(d)}) \\
&= 1 - c(F_1(q_T^{(1)}), \ldots, F_d(q_T^{(d)})) \\
&= 1 - c(1-1/T, \ldots, 1-1/T) \\
&= 1 - \Phi_d\left(\phi^{-1}(1-1/T), \ldots, \phi^{-1}(1-1/T)\right) \\
&= 1/T_R
\end{aligned} \qquad (14)$$

This computation is only based on the dependence between stations, and does not depend on marginal estimations. More precisely, if the probability to compute was $P(\{X^{(1)} \le x^{(1)}\} \cap \ldots \cap \{X^{(d)} \le x^{(d)}\})$, then the uncertainty would have been greater, as it would include correlations and marginal parameter estimation uncertainties.

The preceding equation was applied to the same data set as in section III.1, leading to the results presented in Figure 6. The 90% confidence interval was obtained by bootstrapping (i.e. years are bootstrapped, and the corresponding data of 13 sites are added to the new data set). As an example, the regional return period of the event "observing at least one event of return period ten years" is approximately two years. In the same manner, the regional return period is approximately 15 years for a local return period of 100 years. It can be noticed that the confidence interval for the regional return period does not encompass the two extreme cases of independence ($T_R = \frac{1}{1-(1-1/T)^{13}}$) and total dependence ($T_R = T$).

Moreover, from a manager's point of view, an interesting question may be: for the set of 13 stations, what are the discharges to be protected from in order to obtain a regional return period T_R, if it is assumed that all local protection levels are identical? Figure 6 can be used to compute the local return period T, and the marginal estimations lead to the local quantiles with return period T. As an example, with a regional protection level T_R = 10 years, the local protection level has to be around 70 years. This difference may appear quite large, but once the return period has been translated in terms of discharges thanks to the marginal estimations, this difference between regional and local risks leads to a difference ranging from 25 to 33% between the discharge with local return period 10 years and the discharge to be protected from in order to obtain a regional return period of 10 years. Finally, the preceding computations were processed by using point-estimates of the parameters. Consequently, sampling uncertainty has to be kept in mind when interpreting such results.

III.3. QdF Analysis

The Discharge-Duration-Frequency (QdF) model has been developed in order to generalize the well known Intensity-duration-Frequency (IdF) model used for precipitations analysis [19, 28-30, 44, 50]. It aims at describing the relationship between quantiles computed for mean discharges $V(d,T)$ over a range of durations. More accurately, the model assumes that quantiles decrease as a hyperbolic function of duration:

$$V(d,T) = \frac{V(0,T) - P}{1 + d/\Delta} + P \qquad (15)$$

where P and Δ are parameters which have to be estimated.

One application of the QdF methodology deals with design hydrograph construction. In order to estimate the impact of floods on human activities, hydrologists are asked to provide a hydrograph with a given return period. Of course, this is an almost intractable problem, because a hydrograph can be described in a number of ways, such as peak flow, duration or volume. The QdF approach is used here to compute a Mono-Frequency Synthetic Hydrograph (MFSH), *i.e.* a hydrograph whose mean discharges over a given range of durations have the same return period. The method used to construct the MFSH can be found in Le Clerc [34]. It is based on the study of the hydrograph shapes. When the hypothesis of shape invariance can be accepted, a design hydrograph is computed using the mean of non-dimensional hydrographs, with the discharge divided by peak flow and a synchronization on peak flow. The recession is then corrected in order that each mean discharge over duration d matches with the various quantiles $V(d,T)$ from a QdF analysis. When the various hydrographs have different shapes, Le Clerc [34] proposed dealing with different subsets of hydrographs, related to the various flood origins, e.g. snow melt, thunderstorms or frontal rainfall, or various spatial flood extents within the catchment.

Another approach is based on the multivariate analysis of the mean discharges Vd over different durations d. The hydrograph shape invariance implies a high rank correlation between the mean discharges Vd (the event with strongest peak discharge has also the strongest discharges over durations d_1, d_2,...). On the other hand, if it assumed that the discharges over different durations are not so dependent, the constructed MFSH could then have a smaller probability than that suggested by the return period used for its construction. A Gaussian copula can once again be used in order to estimate this dependency, and to derive what we will call the "multivariate MSFH return period" T_M.

Such an evaluation is described below by two case studies. The first deals with the Zorn river at Waltenheim (688 km²). Mean discharges were computed over durations d=1, 2, 3, 4 and 5 days, and annual maxima were then extracted on 80 years of data. Figure 7a shows the scatterplot matrix of the data: it appears that mean discharges are very strongly correlated over this duration range, which indicates that the shape invariance hypothesis is acceptable.

The Gaussian copula was then estimated with the Phoon method, and its adequacy was checked with the QQ-plot described in section II.3. The multivariate MFSH return period T_M was finally calculated as follows:

$$\begin{aligned}
&\Pr(\text{all mean discharges have a return period greater than T}) = \Pr(\{X^{(1)} > q_T^{(1)}\} \cap ... \cap \{X^{(d)} > q_T^{(d)}\}) \\
&= \Pr(\{F_1(X^{(1)}) > F_1(q_T^{(1)})\} \cap ... \cap \{F_d(X^{(d)}) > F_d(q_T^{(d)})\}) \\
&= \Pr(\{\phi^{-1}(F_1(X^{(1)})) > \phi^{-1}(1-1/T)\} \cap ... \cap \{\phi^{-1}(F_d(X^{(d)})) > \phi^{-1}(1-1/T)\}) \\
&= \bar{\Phi}_d(\phi^{-1}(1-1/T),...,\phi^{-1}(1-1/T)) \\
&= \bar{c}(1-1/T,...,1-1/T) \\
&= 1/T_M \\
&\text{where } \bar{\Phi}_d(y_1,...,y_d) = \Pr(\{Y_1 > y_1\} \cap ... \cap \{Y_d > y_d\}) \text{ with } \boldsymbol{Y} = (Y_1,...,Y_d) \sim N(0,\boldsymbol{\Sigma}) \\
&\text{and } \bar{c}(u_1,...,u_d) = \bar{\Phi}_d(\phi^{-1}(u_1),...,\phi^{-1}(u_d))
\end{aligned} \tag{16}$$

As we are interested in the survival function $\Pr(X_i > x_i \; \forall i = 1,...,d)$, the notations $\bar{c}$ and $\bar{\Phi}_d$ have been introduced in equation (16). However, it should be noticed that there is no simple relationship similar to the equation $\bar{c} = 1 - c$ which holds in the univariate case. Figure 7b illustrates the relationship between the return period used for MFSH construction and the multivariate MFSH return period. It shows that a hydrograph constructed with a 100-years return period has in fact a return period $T_M \approx 160$ years. Conversely, a hydrograph of return period $T_M = 100$ years would be obtained from mean discharges quantiles $V(d,T)$ with a return period T around 65 years. Translated in terms of discharge, this leads to a difference of about 7% between the mean discharge $V(d,T=100 \text{ years})$ and the mean discharge which would lead to a hydrograph return period of 100 years, for the range of durations d considered. The constructed MFSH can thus be considered as suitable for flood scenarios.

The same study was conducted on the Ubaye river at the Lauzet (946 km²), with 45 years of available data. Mean discharges are here computed over durations d = 1, 10, 20 and 30 days, because most floods are related to snow melt, leading to very slow events. The data present a weaker dependence than in the preceding case, especially between the daily duration and the others. Consequently, the ratio between marginal and multivariate return periods T and T_M becomes larger: a hydrograph constructed with T=100 years has in fact a return period T_M about 1000 years (Figure 8a). This result emphasizes the lack of shape invariance for snow-related events: moderate rainfalls can be superimposed on the high baseflow created by snow melt, thus leading to a great variety of shapes at a daily resolution. By contrast, Figure 8b shows the results of the QdF analysis conducted on the same river, but only for events occurring between September and February (rainfall-related floods), with durations d=1, 2, 3 and 4 days. The results are now similar to those obtained with the Zorn River. The construction of a MFSH is thus possible for such flood events.

These two examples provide an illustration of the link between the shape variability of the hydrographs and the dependence between the mean discharges Vd over different durations d. The Gaussian copula provides a quantification of the relationship between the return period T of each mean discharge and the multivariate return period T_M, which can be very different from one catchment to another one. However, the meaning of the multivariate return period T_M is far from obvious, as the sum of the probabilities (p_1+p_2) can be far from one, with p_1=Pr(hydrograph > MFSH) and p_2=Pr(hydrograph < MFSH). A number of intermediate cases can occur, with a cross-over between an observed hydrograph and a MFSH: some mean discharges Vd have a larger return period, and others a smaller, than the prescribed return period used for the MFSH construction. Consequently, such multivariate analysis is more recommended for the study of the hydrograph shape (likelihood of various shape hypotheses) than for an exact assessment of the probability of a design hydrograph.

III.4. Regional Frequency Analysis

Regional frequency analysis is used by hydrologists in order to improve the estimation of high quantiles. The principle is to collect data originating from different locations and to derive a regional distribution of extreme streamflows or rainfalls. Since the precursory work of Dalrymple [11], a number of methodological improvements have been proposed (e.g. [27]). However, a common drawback of almost all regional methods for extremes is that they ignore the spatial dependence of data. This problem has been addressed by Stedinger [58], Hosking and Wallis [26, 27] or Madsen and Rosbjerg [38]. Roughly speaking, these authors found that ignoring intersite dependence led to underestimation of the variance of the estimates, but did not lead to any bias. In the following example, dependence will be explicitly taken into account in the probabilistic model, by means of a Gaussian copula.

Six rainfall stations located around Paris, France, were used (Figure 9). 58 annual maximum values of daily rainfalls were extracted from the years between 1922 and 2003, with 1926, 1936, 1939-1944, 1948-1949, 1952, 1955-1958, 1981-1983, 1991 and 1997-2001 as missing years. The index-flood procedure of Dalrymple [11] comprises an homogenization step, by dividing the at-site samples by the at-site means or medians. For the six study stations, the at-site median values ranged from 28.5 mm to 31.75 mm. Moreover, from a meteorological point of view, the study area can be considered as homogeneous, with little altitudinal range. Consequently, we assume that the six rainfall series arise from an identical GEV distribution. The parameters of this regional distribution are estimated using two models:

In model 1 the intersite dependence is ignored. The multivariate density of a vector of six annual maxima $(x_t^{(1)},...,x_t^{(6)})$ at year *t* can be written as:

$$f_1\,(x_t^{(1)},...,x_t^{(6)}) = \prod_{i=1}^{6} GEV(x_t^{(i)};\alpha,\beta,\xi) \tag{17}$$

In model 2 the intersite dependence is modeled with a Gaussian copula, whose adequacy was checked with the Chi-square plot of section II.3. Using Equation (8), the multivariate density of a vector of six annual maxima $(x_t^{(1)},...,x_t^{(6)})$ at year *t* can be written as:

$$f_2\,(x_t^{(1)},...,x_t^{(6)};\alpha,\beta,\xi,\boldsymbol{\Sigma}) = \left(\prod_{i=1}^{6} GEV(x_t^{(i)};\alpha,\beta,\xi)\right)\left|\boldsymbol{\Sigma}\right|^{-1/2} \times$$

$$\exp\left\{-\frac{\left[\boldsymbol{V}(x_t^{(1)},...,x_t^{(6)};\alpha,\beta,\xi)\right]\left[\boldsymbol{\Sigma}^{-1}-\boldsymbol{I}\right]\left[\boldsymbol{V}(x_t^{(1)},...,x_t^{(6)};\alpha,\beta,\xi)\right]^T}{2}\right\}, \tag{18}$$

with $\boldsymbol{V}(x_t^{(1)},...,x_t^{(6)};\alpha,\beta,\xi) = \left[\phi^{-1}\left(F(x_t^{(1)};\alpha,\beta,\xi)\right),...,\phi^{-1}\left(F(x_t^{(6)};\alpha,\beta,\xi)\right)\right]$

and $F(x;\alpha,\beta,\xi)$ the cdf of the GEV distribution.

The dependence is thus summarized in a correlation matrix Σ. In order to obtain an acceptable number of parameters, the dependence $\rho_{i,j}$ between two stations is assumed to decrease as a function of the distance $d_{i,j}$:

$$\begin{cases} \rho_{i,j} = \gamma_0 \exp\left(-\gamma_1 d_{i,j}\right) \text{ if } i \neq j \\ \rho_{i,i} = 1 \end{cases} \tag{19}$$

Such a parameterization can be compared to classical variograms used in geostatistics. Because of the complexity of this model, a Bayesian estimation scheme using MCMC algorithms is applied. Some details of these methods in a hydrological context can be found for instance in the papers by Perreault et al. [47, 48], Thyer et al. [61], Marshall et al. [41] and

Renard et al. [53]. The likelihood of the data is simply computed from the product of the multivariate densities of observations, *i.e.* for one of the models *k=1,2* described above:

$$p_k(\boldsymbol{X} \mid \boldsymbol{\theta_k}) = \prod_{t=1}^{n} f_k\left(x_t^{(1)},...,x_t^{(6)};\boldsymbol{\theta_k}\right) \tag{20}$$

with $\boldsymbol{\theta_1} = (\alpha, \beta, \xi)$ and $\boldsymbol{\theta_2} = (\alpha, \beta, \xi, \boldsymbol{\Sigma}) = (\alpha, \beta, \xi, \gamma_0, \gamma_1)$.

The prior distribution of the shape parameter ξ is chosen to be a Gaussian distribution with zero mean and standard deviation 0.3, which implies that the interval [-0.6;0.6] encompasses more than 95% of the density. This prior distribution can be compared with the Martins and Stedinger [42] geophysical prior, which is less variable and is entirely included in the interval [-0.5;0.5]. For other parameters, almost non-informative priors are set, by using uniform distributions with large variances, namely $\alpha \sim U[0,1000]$, $\beta \sim U[-10000,10000]$, $\gamma_0 \sim U[0,1]$ and $\gamma_1 \sim U[0,10]$. Finally, the independence between these prior marginal distributions is assumed, in order to derive the multivariate prior distribution $\pi_k(\boldsymbol{\theta_k})$ of the parameters of the model *k*.

The posterior distribution is finally obtained up to a constant of proportionality by:

$$p_k(\boldsymbol{\theta_k} \mid \boldsymbol{X}) \propto p_k(\boldsymbol{X} \mid \boldsymbol{\theta_k})\pi_k(\boldsymbol{\theta_k}) \tag{21}$$

Samples arising from the posterior distribution of the parameters are generated with MCMC algorithms in the following way. First of all, at-site maximum likelihood estimations are used to derive a rough estimate of the posterior means and variances of the parameters α, β and ξ. Because almost non-informative priors are used, the posterior distribution will be mostly influenced by the likelihood, thus making the ML-estimates relevant. For the two remaining parameters of model 2, the exponential model described in equation (19) is fitted to the empirical correlations estimated by the method of Phoon, using a classical least-square approach. As a second step, these estimations are used as starting parameters of a Metropolis algorithm (see Renard et al. [53] for a detailed description), which was run for 100 000 iterations, and whose convergence was checked using the approach suggested by Gelman et al. [20]. Finally, the last 50 000 iterations are used to perform the inference.

Figure 10 describes the relationship between distance and dependence. The solid line represents the median exponential decrease described in Equation (19), and the dashed lines denote a 90% posterior confidence interval. The model of dependence/distance relationship seems acceptable for the six series studied. Figure 11 compares the posterior distributions of parameters α, β and ξ obtained with the two models. For the first two parameters, the results are almost identical for the mean posterior values, but the estimates of model 2, which takes into account the intersite dependence, have a larger variance. This is consistent with the conclusions of Stedinger [58], Hosking and Wallis [26, 27] and Madsen and Rosbjerg [38] described earlier. Conversely, the posterior variances of the shape parameter are almost identical for the two models, with a slight shift between the two distributions. This result shows that the effect of intersite dependence is not necessarily the same for all parameters. The implications for quantiles estimates are shown in Figure 12: the difference between the two models, in terms of posterior variance, is stronger between moderate quantiles (i.e. 10-years return period) than between high quantiles (i.e. 100-years return period).

These results also have implications for regional methods of extrapolation. As an illustration, the FORGEX method [16, 51, 59] uses the series of the annual maximum of the standardized values observed over a measurement network (netmax series) in order to extrapolate the distribution of a target site. This method uses the following observation [52]: if the *N* stations of the network are independent, then the netmax variable *Y* has the following distribution:

$$
\begin{aligned}
F_Y(y) &= \Pr(Y \le y) \\
&= \Pr\left(Max\left(X_1,...,X_N\right) \le y\right) \\
&= \Pr\left(\left\{X_1 \le y\right\} \cap ... \cap \left\{X_N \le y\right\}\right) \\
&= \prod_{i=1}^{N} \Pr\left(X_i \le y\right) \\
&= \left(F_X(y)\right)^N
\end{aligned}
\tag{22}
$$

In other terms:

$$
\begin{aligned}
-\log(-\log(F_Y(y))) &= -\log\left(-\log\left(\left(F_X(y)\right)^N\right)\right) \\
&= -\log(N) - \log(-\log(F_X(y)))
\end{aligned}
\quad , \tag{23}
$$

so that in a Gumbel repair, the distribution of the netmax lies at a distance of log(N) to the left of the population growth curve (*i.e.* the distribution of the standardized values). In order to take into account the effect of intersite dependence, this distance is in fact replaced by log(N_e), where N_e is an equivalent number of stations. The latter concept has been used for a long time [43] in order to take into account the information redundancy caused by dependence. However, in the present situation, this approach suffers from two drawbacks, whatever the method used to estimate the equivalent number of stations. First of all, the preceding results show that the effect of intersite dependence is not identical for all parameters, or equivalently for the whole quantile curve. Consequently, using a unique value as N_e to summarize the effect of dependence is questionable. Secondly, the parallelism between at-site and netmax growth curves is only true under the hypothesis of independence. As an illustration, a simple bidimensional case is explored: for a network of two sites, whose dependence is described with a Gaussian copula with correlation 0.8, the at-site distribution (a GEV(1,0,-0.5)) is plotted in a Gumbel repair, together with the netmax distribution under the independence hypothesis and the "true" Gaussian copula hypothesis. Figure 13 (left panel) shows that the at-site and the netmax distributions are parallel assuming independence, but not with the prescribed model of dependence. In the right panel, the gap between these two distributions is shown: in the independence case, the gap is constant and equal to log(2), but this not the case in the Gaussian copula case, where this gap depends on marginal values, and can therefore not be equal to log(N_e), whatever the N_e used. The same computation is made for the preceding case study (Figure 14), with identical conclusions. In this case, the distributions are plotted using point-estimates of the parameters. Consequently, they are affected by sampling uncertainty, which is not shown in the figure for clarity, but which is far from negligible, thus complicating the conclusions of the comparison.

The case study could be enhanced in several ways. First of all, the assumption of a single shared distribution may be relaxed. Rather than dividing the at-site samples by an index variable, the spatial variability of the parameters could be modeled with covariables such as altitude or distance from the sea. Such an approach has been proposed by Diggle et al. [14], Cooley [9] and Cooley et al. [10]. Moreover, from a Bayesian perspective, the use of a proper prior distribution could improve the accuracy of estimates. Bayesian model checking or Bayesian comparison of models could also be useful for deriving a more complete validation of the model assumptions. Finally, it is clear that the preceding conclusions are dependent on the Gaussian copula hypothesis, which remains questionable. Further investigations are needed to refine regional frequency analysis, using more complex dependence structures.

III.5. The Gaussian copula can fail

The need to check the adequacy of the copula before using it for any computation has been pointed out in section II. The following case study illustrates the fact that a Gaussian copula is far from being a universal tool. 55 flood events were selected from the daily discharge series of the Ubaye river at Barcelonette (549 km²; 1904-2001). The scatterplot of the peak discharge of the hydrograph versus the volume computed above a threshold equal to the half of the peak discharge is shown in Figure 15a. At first sight, the dependence structure of the data seems complex. Nevertheless, the shape of the scatterplot may be due to the marginal structure of the observations: on this river, flood events can result from snow melt or heavy rainfalls. Both marginal distributions are therefore likely to derive from a population mixture model. In order to erase the influence of marginal distributions, a normal score transformation is applied to the data, leading to the scatterplot shown in Figure 15c. Snow-related events are denoted by crosses, and rainfall-related ones by triangles. Despite this transformation, the two types of events still lead to different dependence structures: for snow-related events, peak flows and volumes are more correlated than for rainfall-related ones. It is thus clear that modeling the dependence structure with a single correlation parameter will lead to very poor results. As an illustration, the contour of the estimated bivariate Gaussian density is also plotted in the Figure 15c. It clearly appears that the distribution of the points is not consistent with a Gaussian description. This departure from normality can also be viewed in the Chi-square plot of Figure 15d, where high quantiles have a tendency to deviate from the theoretical line $y=x$.

A more subtle problem arises from the asymptotic properties of the copula, also known as the tail dependence properties. More accurately, the Gaussian copula implies the asymptotic independence of marginal values, which means that $\Pr\left(F_X(X) > p \mid F_Y(Y) > p\right) \xrightarrow[p\to 1]{} 0$. In other words, the dependence weakens for very extreme events. This property has a strong influence for the computation of very low probabilities, concerning events lying outside the range of observations. More explicitly, the risk can be strongly underestimated if an asymptotically independent model is used with asymptotically dependent data. For this reason, empirical or physical evidence of tail independence is necessary before using such a copula for extrapolation. Coles et al. [7] provide a number of tools for exploring the asymptotic properties of a multivariate data set.

This phenomenon will be illustrated with a simulated case study. The following variables are considered:

$$X \sim N(0;1)$$

$$Y = X + Z, \text{ where } Z \mid X \sim N\left(0; \frac{\exp(-X^2/4)}{4}\right). \tag{24}$$

The idea behind this construction is to create a couple of variables whose dependence increases with marginal values. A sample of (X,Y) with size 10^6 was thus generated. In order to deal with known marginal distributions, data were finally transformed to fit a GEV distribution with parameters (1,0,-0.5):

$$(\tilde{X}, \tilde{Y}) = \left(G^{-1}\left\{\hat{F}_X(X)\right\}, G^{-1}\left\{\hat{F}_Y(Y)\right\}\right) \tag{25}$$

where G is the cdf of the GEV distribution and $\hat{F}_X$ is the empirical cdf of X.

The first 100 couples were used for the copula estimation. The scatterplot of these data is shown in Figure 16a. In order to focus on the problems caused by the dependence structure, the marginal parameters were not estimated but fixed at the true value (1,0,-0.5). Figure 16b shows the scatterplot of the normalized data, and the contour of the bivariate density

estimated with the Phoon method. At first sight, the Gaussian copula seems to be a reasonable description of the dependence of the data.
The Gaussian copula is now used to extrapolate far outside the observations range. More accurately, the probability $p = \Pr(\tilde{X} > u, \tilde{Y} > u)$ is computed, with u possibly lying outside the observations range. This probability is estimated with the Gaussian copula as $p = \bar{\Phi}_d\left(\phi^{-1}(G(u)), \phi^{-1}(G(u))\right)$, and is compared to the theoretical value, empirically estimated with the total sample of length 10^6. The results are shown in Figure 17, where the probability described above is expressed in terms of return period (T=1/p). On the left-hand side, it can be observed that the Gaussian copula is adequate in the observation range. Conversely, the quantile curves differ strongly in the extrapolation range: the curve associated with the Gaussian copula grows faster than the theoretical one, which leads to a severe underestimation of risk. As an example, the quantile is under-estimated by about 38% for a return period of 10^3 years, and by about 58% for a return period of 10^4 years. This can be explained by the fact that the simulated data are not asymptotically independent. The probability of observing an extreme value in both components is thus higher than that suggested by the Gaussian copula model.

IV. Discussion

The case studies presented in sections III.1 to III.4 demonstrate the usefulness and the relative simplicity of the Gaussian copula for dealing with multivariate events. Conversely, section III.5 highlights its limitations and emphasizes the error risk related to extrapolation. As no theoretical argument can be employed to justify the use of this copula, this raises the question of how to decide, in practical cases, if this model is acceptable or should be avoided. Two different aspects have to be explored for this purpose. The first is to check that the Gaussian copula is consistent with the observed data. This implies that both the marginal distributions and the dependence structure have to be in agreement with the observations. For the marginal distributions, a number of standard univariate tools can be used (e.g. goodness-of-fit tests or QQ-plots). Moreover, univariate extreme value theory provides a strong theoretical justification for the use of extreme value distributions. After marginal transformations, the distribution is determined by a simple multivariate Gaussian model. Diagnoses for multivariate normality have also been extensively studied, thus providing a number of tools for model checking [40].
The second aspect is more difficult to handle, and is related to the asymptotic dependence properties of the data. From a theoretical point of view, it can be shown that the limit distribution of the componentwise maxima (Equation (2)) has the property of asymptotic dependence, except in the case of exact independence. Consequently, the Gaussian copula might be considered at first sight to be inadequate for describing these data, because it is asymptotically independent. Alternative multivariate distributions may be used [8], *e.g.* using the logistic family. Such a development has been undertaken by Tawn [60] and Schalter and Tawn [55] for an arbitrary number of dimensions, and may be applied with additional hypotheses in order to reduce the number of parameters (2^d-1). However, this limitation has to be mitigated, because the limit distribution of Equation (2) is valid when the block size tends toward infinity. In most hydrological applications, the block size is typically of 365 days, with non-independent daily data, and a number of zero values in the case of rainfall. It is thus possible that the asymptotic arguments used to derive the limit distribution of componentwise maxima do not hold in real-life hydrological studies. Although the latter argument can also be argued for univariate extreme value distributions, it seems that the problem is more critical in

the multidimensional case. For instance, Bortot et al. [3] reported empirical evidence of asymptotically independent data in the oceanographic field. Alternatively, Hosking and Wallis [26] argued that a Gaussian copula model was fairly well supported by British annual flood series, although in some cases, stronger dependence seemed to occur at higher levels. From a practical point of view, a given data set of componentwise maxima will generally not include enough data to provide empirical evidence of asymptotic dependence or independence. Physical arguments may be used to discriminate between these two hypotheses, by using for example meteorological model simulations, or by exploiting the knowledge of historical extreme events.
Consequently, our point of view is that the Gaussian copula should not be used for the computation of very low probabilities, unless strong arguments can be derived to justify the asymptotic properties of the data. Conversely, the Gaussian copula can be very useful within the observations range. As an illustration, in the case study in section III.4, the copula can be used to take into account intersite dependence in regional frequency analysis, and to derive a more realistic quantification of the uncertainties than with the independence hypothesis, which is in any case a very crude model for describing dependent data.

V. Conclusion and perspectives

The aim of this article was to present some possible applications of the Gaussian copula in the flood mitigation field. The main advantage of this method is its simplicity, even with more than two or three dimensions. Rough estimates can thus be obtained in problems involving an evaluation of the dependence of variables. Moreover, the Gaussian copula can be used to simulate values with prescribed correlations and marginal distributions. Four case studies were used to demonstrate its usefulness in the contexts of field significance determination, regional risk analysis, QdF models with design hydrograph derivation and regional frequency analysis. Nevertheless, this tool is far from universal, as it is based on no theoretical justification. The suitability of the Gaussian dependence model therefore has to be properly checked. Moreover, even if the copula seems well fitted to observed values, caution is needed before using it for extrapolation.
Alternative copula families can be used in order to improve the dependence model arising from the Gaussian copula. For instance, the Student copula may be more appropriate if tail dependence is observed in the data. Nevertheless, the choice of the more suitable copula for modeling a data set is not straightforward. Various criteria may be used for this purpose, but they will only reflect a particular feature of the data. For instance, two models may lead to an adequate fit to the data on the basis of a given criterion, but to contrasted results in extrapolation. The choice of a copula in such a case is problematic, especially if the data set is not informative enough to provide relevant indications about the asymptotic dependence properties. Modeling uncertainty can thus be an important part of the overall uncertainty. Another problem concerns the parameter estimation uncertainties, because the maximum likelihood method can usually not be used, due to the great number of parameters to be estimated. Computations based on the Fisher information matrix are thus impossible. Finally, whatever the copula used, the lack of theoretical justification remains a limitation for extrapolation.
The preceding approaches only consist in modeling the extreme part of the variables distribution. An alternative consists in studying the whole distribution, and evaluating how dependence evolves for high values. For instance, Bortot et al. [3] applied a model where the tail dependence is described using a multivariate Gaussian distribution, after suitable marginal transformations. In spirit, this model is similar to that of the Gaussian copula. The major difference is that the data are not block maxima, and that the Gaussian tail model is only

applied above a multivariate threshold $(u_1,...,u_d)$, which has to be high enough to ensure that the asymptotic assumptions hold, both for marginal and joint distributions. Moreover, given that the distribution belongs to the domain of attraction of a multivariate extreme value distribution, some theoretical results related to the theory of regular variation [2] can be used in order to compute the probability of extreme sets. An example of such study is provided by De Haan and De Ronde [12]. Such an approach is appealing, because it is physically more convincing: couples of values are describing the same physical event, while componentwise maxima lead to the use of couples of values which may describe two distinct events. Nevertheless, some difficulties still hold. Firstly, obtaining a sample of independent values is not obvious. As an example, if the data consist of daily discharges at two locations, a sampling strategy has to be found which ensures the independence of successive values and preserve enough data for estimation. Secondly, this approach can only be used to compute the probability of sets for which both components are extreme. This can be a strong limitation: in the confluence problem, as an example, an extreme downstream discharge can result from a high value in only one of the two upstream flows. Heffernan and Tawn [24] recently proposed a semi-parametric conditional approach which overcomes this difficulty.
More generally, we believe that multivariate extreme events analysis is likely to improve hydrological risks assessment, as emphasized by Katz et al. [31]. In a number of situations, multivariate events are used without a formal statistical model to account for dependence. As an example, the well-known index flood methodology [11] is intended to improve quantiles estimations by using data from several sites. Unfortunately, among other problems, ignoring spatial dependence leads to an underestimate of the quantiles uncertainty. Alternatively, most hydrological extreme events (floods or droughts) are intrinsically multivariate, as noted by Adamson et al. [1]. Potential damage is thus likely to be a function of several random variables. Hydrologists involved in risk assessment should thus be attentive with the progress achieved by statisticians in multivariate extremes theory.

Acknowledgements

The financial support provided by Cemagref and EDF for the PhD research of B. Renard is gratefully acknowledged. Discharge data series are derived from the national HYDRO database from the French Ministry of Environment, and the rainfall series have been reviewed by Météo-France within the European IMFREX project. Both organisms are acknowledged. We also thank two anonymous reviewers for their helpful comments.

VI. Bibliography

[1] Adamson PT, Metcalfe AV, Parmentier B. Bivariate extreme value distributions: an application of the Gibbs sampler to the analysis of floods. Water Resources Research 1999; 35:2825-2832.
[2] Bingham NH, Goldie CM, Teugels JL. Regular variation. New York: Cambridge University Press, 1987.
[3] Bortot P, Coles S, Tawn JA. The multivariate Gaussian tail model: an application to oceanographic data. Appl. Statis. 2000; 49:31-49.
[4] Caperaa P, Fougeres AL, Genest C. Bivariate distributions with given extreme value attractor. J. Multivar. Anal. 2000; 72:30-49.
[5] Cherubuni U, Luciano E, Vecchiato W. Copula Methods in Finance: Wiley, 2004.
[6] Coles S. An Introduction to Statistical Modeling of Extreme Values. London: Springer-Verlag, 2001.
[7] Coles S, Heffernan JE, Tawn JA. Dependence measures for extreme value analyses. Extremes 1999; 2:339-365.
[8] Coles S, Tawn JA. Modelling extreme multivariate events. J. R. Stat. Soc. Ser. B-Methodol. 1991; 53:377-392.
[9] Cooley D. Statistical Analysis of Extremes Motivated by Weather and Climate Studies: Applied and Theoretical Advances, PhD thesis, University of Colorado, 2005
[10] Cooley D, Nychka D, Naveau P. A Spatial Bayesian Hierarchical Model for a Precipitation Return Levels Map. In: Extreme Value Analysis. Gothenburg, Sweden, 2005.
[11] Dalrymple T. Flood frequency analyses. In: Water-supply paper 1543-A: US Geological Survey, 1960.
[12] De Haan L, De Ronde J. Sea and Wind: Multivariate Extremes at Work. Extremes 1998; 1:7-45.
[13] De Michele C, Salvadori G, Canossi M, Petaccia A, Rosso R. Bivariate statistical approach to check adequacy of dam spillway. J. Hydrol. Eng. 2005; 10:50-57.
[14] Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. J. R. Stat. Soc. Ser. C-Appl. Stat. 1998; 47:299-326.
[15] Douglas EM, Vogel RM, Kroll CN. Trends in floods and low flows in the United States: impact of spatial correlation. J. Hydrol. 2000; 240:90-105.
[16] Faulkner DS, Jones DA. The FORGEX method of rainfall growth estimation - III: Examples and confidence intervals. Hydrol. Earth Syst. Sci. 1999; 3:205-212.
[17] Favre AC, El Adlouni S, Perreault L, Thiemonge N, Bobee B. Multivariate hydrological frequency analysis using copulas. Water Resources Research 2004; 40.
[18] Fisher RA, Tippett LH. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Cambridge Phil. Soc. 1928; 24.
[19] Galéa G, Prudhomme C. Notions de base et concepts utiles pour la compréhension de la modélisation synthétique des régimes de crue des bassins versants au sens des modèles QdF. Rev. Sci. Eau 1997; 1:83-101.
[20] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis: Chapman & Hall, 1995.
[21] Genest C, McKay J. The joy of copulas: bivariate distributions with uniform marginals. The American Statistician 1986; 40:280-283.

[22] Genest C, Rivest L-P. A characterization of Gumbel's family of extreme value distributions. Statistics and Probability Letters 1989; 8:207-211.
[23] Grimaldi S, Serinaldi F. Asymmetric copula in multivariate flood frequency analysis. Adv. Water Resour. 2006; 29:1155-1167.
[24] Heffernan JE, Tawn JA. A conditional approach for multivariate extreme values. Journal of the Royal Statistical Society 2004; 66:497-546.
[25] Herr HD, Krzysztofowicz R. Generic probability distribution of rainfall in space: the bivariate model. J. Hydrol. 2005; 306:234-263.
[26] Hosking JRM, Wallis JR. The effect of intersite dependence on regional flood frequency analysis. Water Resources Research 1988; 24:588-600.
[27] Hosking JRM, Wallis JR. Regional Frequency Analysis: an approach based on L-Moments. Cambridge, UK: Cambridge University Press, 1997.
[28] Javelle P. Caractérisation du régime des crues: le modèle débit-durée-fréquence convergent. Approche locale et régionale, PhD thesis, INP Grenoble, Cemagref Lyon, 2001
[29] Javelle P, Grésillon JM, Galéa G. Discharge-duration-Frequency curves modeling for floods and scale invariance. Comptes Rendus de l'Académie des Sciences, Sciences de la terre et des planètes. 1999; 329:39-44.
[30] Javelle P, Ouarda T, Lang M, Bobee B, Galéa G, Grésillon JM. Development of regional flood-duration-frequency curves based on the index-flood method. J. Hydrol. 2002; 258:249-259.
[31] Katz RW, Parlange MB, Naveau P. Statistics of extremes in hydrology. Adv. Water Resour. 2002; 25:1287-1304.
[32] Kelly KS, Krzysztofowicz R. A bivariate meta-Gaussian density for use in hydrology. Stoch. Hydrol. Hydraul. 1997; 11:17-31.
[33] Kendall MG. Rank correlation methods. London: Griffin, 1975.
[34] Le Clerc S. Revisiter la notion de scénario hydrologique de référence pour la caractérisation des inondations, PhD thesis, Université Joseph Fourier Grenoble, Cemagref Lyon, 2004
[35] Le Clerc S, Lang M. Flood frequency analysis downstream confluences. Comparison between bivariate densities and experimental data. In: International conference on flood estimation. Berne, Switzerland, 2002; 295-304.
[36] Lettenmaier DP, Wood EF, Wallis JR. Hydro-Climatological Trends in the Continental United-States, 1948-88. J. Climate 1994; 7:586-607.
[37] Livezey RE, Chen WY. Statistical field significance and its determination by Monte Carlo techniques. Monthly Weather Review 1983; 111:46-59.
[38] Madsen H, Rosbjerg D. The partial duration series method in regional index-flood modeling. Water Resources Research 1997; 33:737-746.
[39] Mann HB. Nonparametric tests against trend. Econometrica 1945; 13:245-259.
[40] Mardia KV. Tests of univariate and multivariate normality. In: Krishnaiah PR, ed. Handbook of Statistics 1: analysis of variance. Amsterdam, Holland, 1980; 279-320.
[41] Marshall L, Nott D, Sharma A. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resources Research 2004; 40.
[42] Martins ES, Stedinger JR. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. Water Resources Research 2000; 36:737-744.
[43] Matalas NC, Langbein WB. Information content on the mean. Journal of Geophysical Research 1962; 67:3441-3448.
[44] Meunier M. Regional flow-duration-frequency model for tropical island of Martinique. J. Hydrol. 2001; 247:31-53.

[45] Mikosch T. How to model multivariate extremes if one must? Stat. Neerl. 2005; 59:324-338.
[46] Mousavi NS. Composition des lois élémentaires en hydrologie régionale : application à l'étude des régimes de crue, PhD thesis, Université Joseph Fourier Grenoble, Cemagref Lyon, 1997
[47] Perreault L, Bernier J, Bobee B, Parent E. Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. J. Hydrol. 2000; 235:221-241.
[48] Perreault L, Bernier J, Bobee B, Parent E. Bayesian change-point analysis in hydrometeorological time series. Part 2. Comparison of change-point models and forecasting. J. Hydrol. 2000; 235:242-263.
[49] Phoon KK, Quek ST, Huang HW. Simulation of non-Gaussian processes using fractile correlation. Probab. Eng. Eng. Mech. 2004; 19:287-292.
[50] Prudhomme C. Modèles synthétiques des connaissances en hydrologie, PhD thesis, Université Montpellier II, Cemagref Lyon, 1995
[51] Reed DW, Faulkner DS, Stewart EJ. The FORGEX method of rainfall growth estimation - II: Description. Hydrol. Earth Syst. Sci. 1999; 3:197-203.
[52] Reed DW, Stewart EJ. Inter-site and inter-duration dependence in rainfall extremes. In: Turkman VBaKF, ed. Statistics for the Environment 2 : Water related issues. Chichester, UK: Wiley, 1994; 125-143.
[53] Renard B, Garreta V, Lang M. An application of Bayesian analysis and MCMC methods to the estimation of a regional trend in annual maxima. submitted to Water Resources Research. 2006.
[54] Salvadori G, De Michele C. Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. Water Resources Research 2004; 40.
[55] Schalther M, Tawn JA. A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika 2003; 90:139-156.
[56] Sklar A. Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Stat. Univ. Paris 1959; 8:229-231.
[57] Song PXK. Multivariate dispersion models generated from Gaussian copula. Scand. J. Stat. 2000; 27:305-320.
[58] Stedinger JR. Estimating a regional flood frequency distribution. Water Resources Research 1983; 19:503-510.
[59] Stewart EJ, Reed DW, Faulkner DS, Reynard NS. The FORGEX method of rainfall growth estimation - I: Review of requirement. Hydrol. Earth Syst. Sci. 1999; 3:187-195.
[60] Tawn JA. Modelling multivariate extreme value distributions. Biometrika 1990; 77:245-253.
[61] Thyer M, Kuczera G, Wang QJ. Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation. J. Hydrol. 2002; 265:246-257.
[62] Troutman BM, Karlinger MR. Regional flood probabilities. Water Resources Research 2003; 39.
[63] Yue S, Wang CY. Regional streamflow trend detection with consideration of both temporal and spatial correlation. Int. J. Climatol. 2002; 22:933-946.

LIST OF CAPTIONS

TABLES

FIGURES

by the estimated empirical cdf. (c) Normalized scatterplot, with estimated Gaussian distribution. Snow-related events are denoted by crosses, and rainfall-related ones by triangles. (d) Chi-square plot.

Figure 16. Scatterplot of the first 100 simulated values. (a): raw values, (b): normalized values and isolines of the estimated bivariate Gaussian density.

Figure 17. Return period associated with the probability $P(\tilde{X} > u, \tilde{Y} > u)$. Solid line represents the theoretical probability, and dotted line the value estimated with the Gaussian copula. (a) u is in the observations range, (b) extrapolation.

Local α	Observed number of significant tests	Critical number of significant tests at regional risk		
		1%	5%	10%
0.01	3	3 *	2 *	2 *
0.05	5	6	4 *	3 *
0.1	5	8	5 *	4 *

Table 1

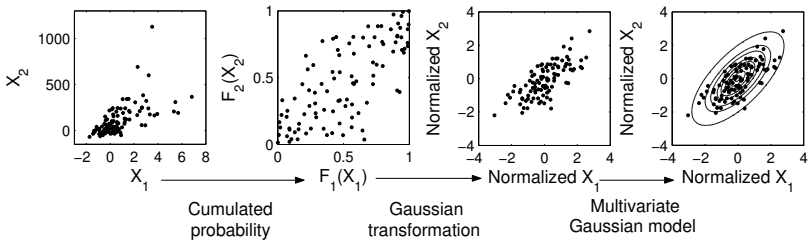

Figure 1

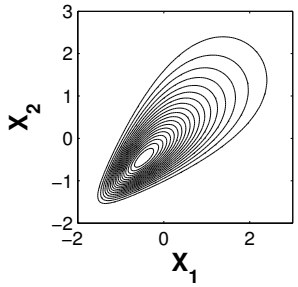

Figure 2

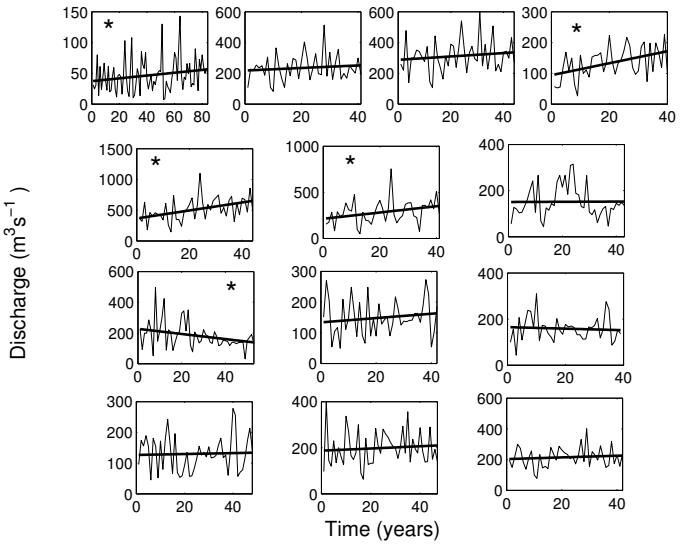

Figure 3

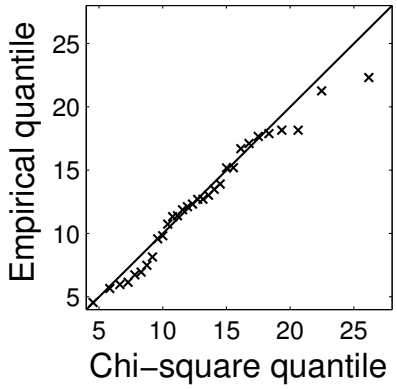

Figure 4

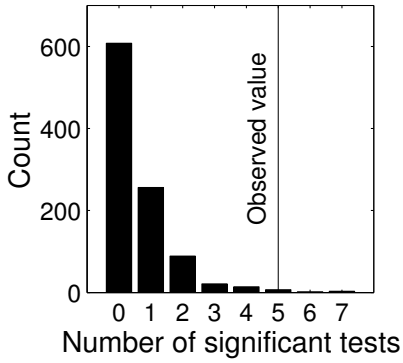

Figure 5

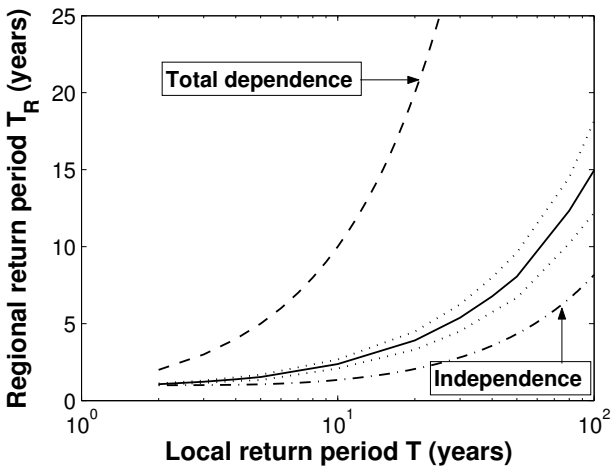

Figure 6

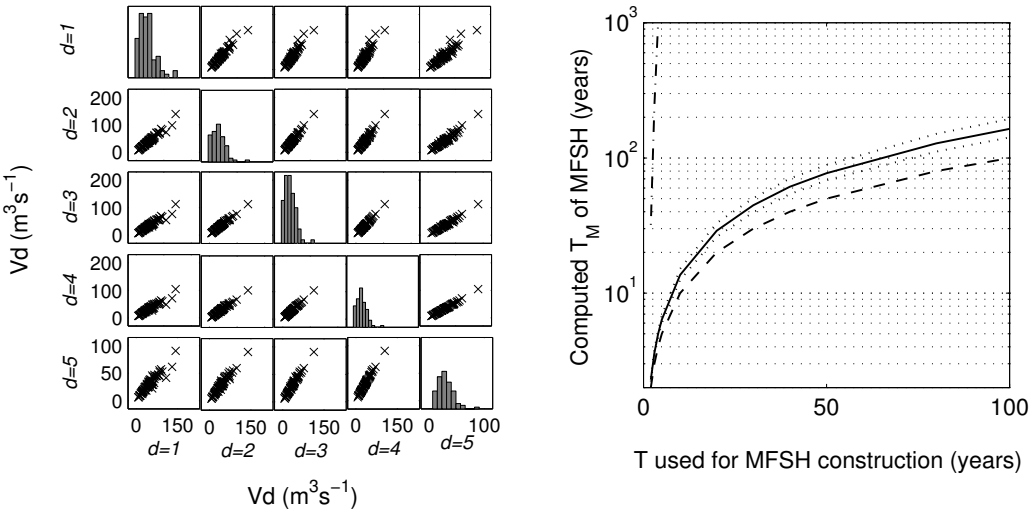

Figure 7

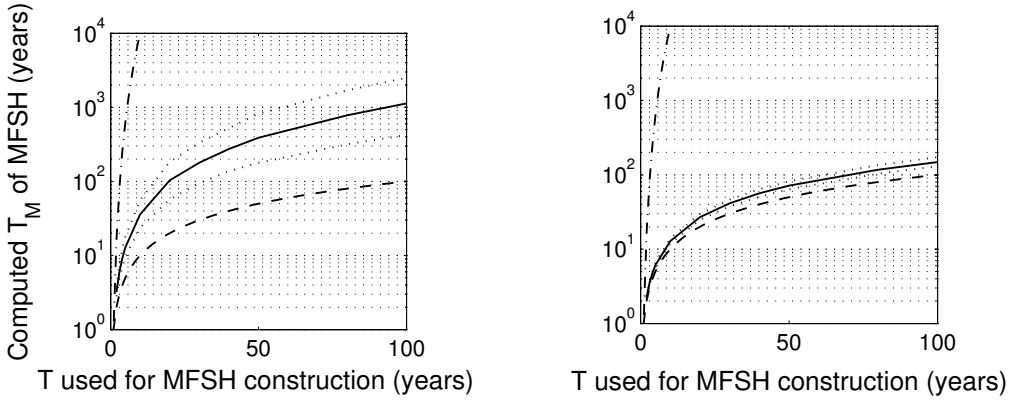

Figure 8

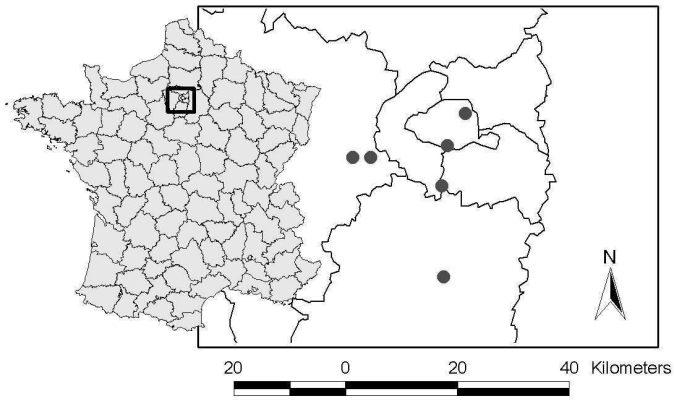

Figure 9

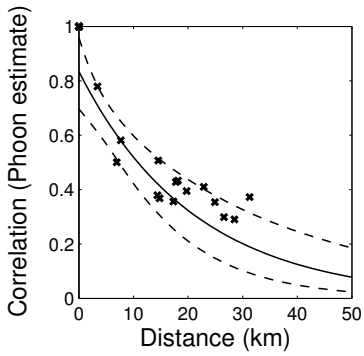

Figure 10

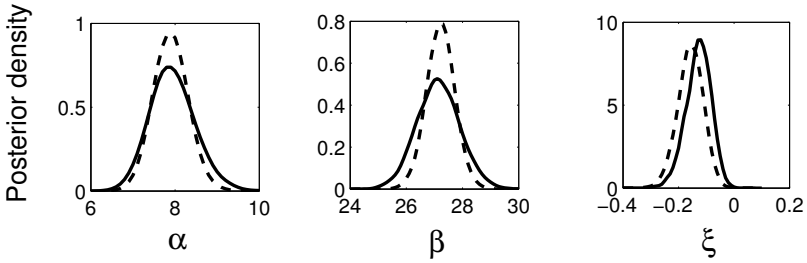

Figure 11

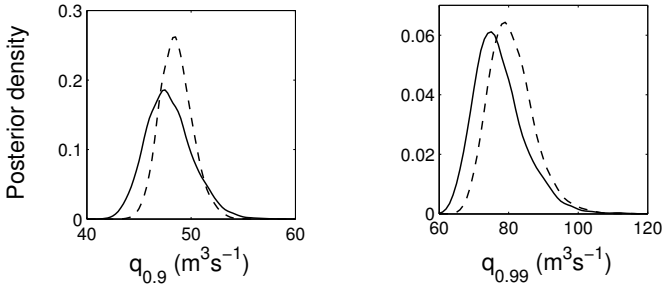

Figure 12

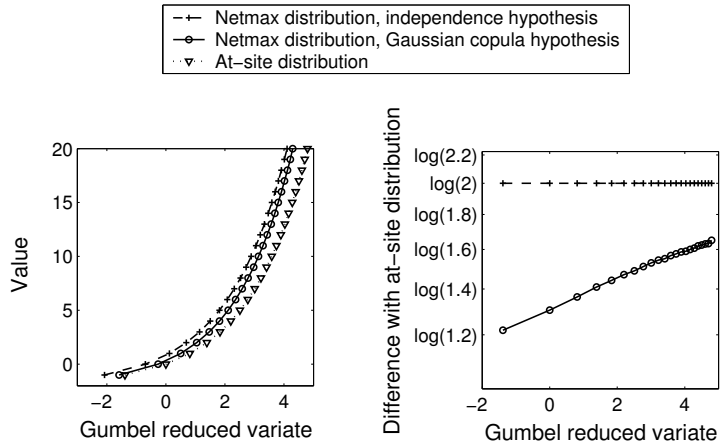

Figure 13

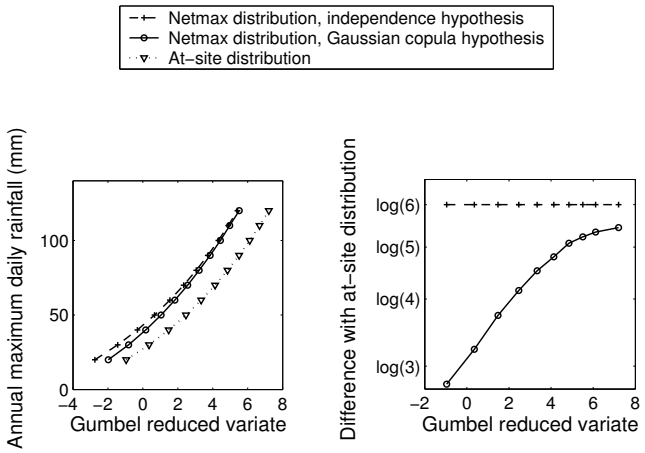

Figure 14

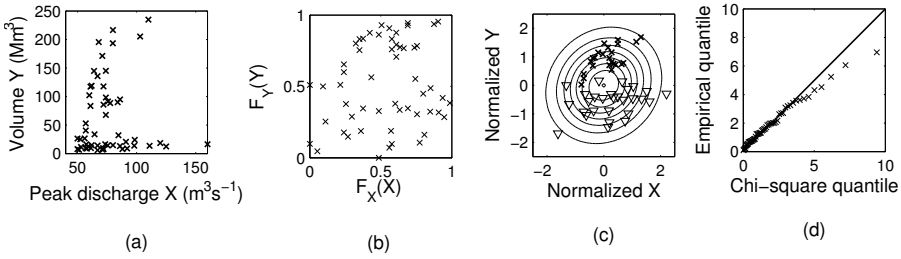

Figure 15

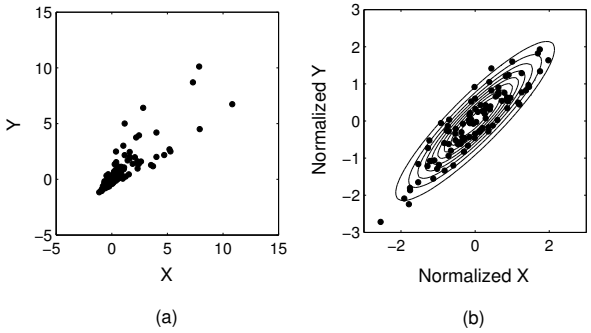

Figure 16

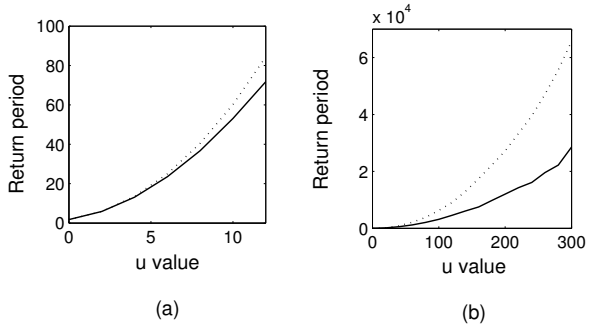

Figure 17