Modeling the statistical dependence in hydrology using “Copulas”

Carlo De Michele

DIIAR, Politecnico di Milano
Abstract
In many hydrological problems several are the (random) variables that play a significant role in the modelization of the phenomenon under investigation, and such variates are generally not independent. For instance, different combinations of rainfall intensity and storm duration may generate storms showing quite different characteristics, and the river management may strongly depend upon the joint features of flood peak and flood volume. Therefore, it is often of fundamental importance to be able to relate the marginal distributions of different variables in order to obtain a joint law describing the main features of the observed hydrological events.

Recent advances in applied mathematics have shown that “Copulas” may represent an useful tool for investigating the statistical behaviour of dependent variables. Practically, “Copulas” are operators on the family of (one-dimensional) probability distributions which generate multivariate laws with specified properties. In fact, given two (continuous) random variables X and Y, with marginal distributions $F_X$ and $F_Y$, there exists a correspondence between their joint law $F_{XY}$ and a proper 2-Copula $C$, Sklar's Theorem. The interesting point is that the properties of $F_{XY}$ can be discussed in terms of the structure of $C$: in fact, it is precisely the copula which captures many of the features of a joint distribution, and dependence properties and measures of association between X and Y can also be investigated in terms of copulas. A further important concept is that of concordance — intuitively, X and Y are concordant if “large” values of one variable tend to be associated with “large” values of the other, and “small” values of one with “small” values of the other; also concordance properties can easily be discussed in terms of 2-Copulas.

In this talk, we shall outline the general mathematical framework of copulas, discussing their usefulness in hydrological problems and stressing several possibilities of application.
Modeling the statistical dependence of the following variables

- Storm rainfall intensity - Storm duration
- Flood peak - Flood volume
- Flood peaks at different river sites
In the presentation:

i) An overview of “Copulas”

ii) Analysis of some bivariate and multivariate problems in hydrology

iii) Case studies

iv) Conclusions
An overview of “Copulas”

Sklar’s Theorem

Let consider 2 random variables $X$ and $Y$ with marginal distributions $F_X(x)$ and $F_Y(y)$.
Let denote $F_{XY}(x,y)$ the joint cumulative distribution.

Exists a unique 2-copula $C$ that

$$F_{XY}(x,y) = C(F_X(x), F_Y(y))$$

- The properties of $F_{XY}$ can be investigated and discussed in terms of the dependence function $C$.

- The copula describes and models the dependence structure between the random variables independently of the marginal laws of the variables involved.
An overview of “Copulas”

Similar considerations are possible for \( N \)-variate problems.
Let consider \( N \) random variables, \( X_1, X_2, \ldots, X_N \) with marginal distributions \( F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_N}(x_N) \).
Let denote \( F_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) \) the joint cumulative distribution.

Exists a unique \( N \)-copula \( C \) that

\[
F_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) = C(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_N}(x_N)).
\]
An overview of “Copulas”

In statistical Literature, several families of “Copulas” are developed:

Bivariate Type-I Extreme Values
(Johnson & Kotz, 1972)
Logistic model

\[
C_\delta(u, v) = e^{-\left((-\ln u)^\delta + (-\ln v)^\delta\right)/\delta}
\]

dependence parameter \(\delta \geq 1\).

The model considers only positive dependence

\(\delta = 2\)
An overview of “Copulas”

Frank’s copula
(Frank, 1972)

\[ C_\delta(u, v) = -\frac{1}{\delta} \ln \left\{ \eta - \frac{(1 - e^{-\delta u})(1 - e^{-\delta v})}{\eta} \right\} \]

where \( \eta = 1 - e^{-\delta} \)

\( \delta \) is variable in the range \(-\infty < \delta < \infty\)
\( \delta = 0 \) corresponds to independence of the two variables.
It models positive and negative dependence
An overview of “Copulas”

The dependence parameter $\delta$ can be linked to some “measures of association”, e.g., *Kendall’s tau*, $\tau$

$$
\tau = \Pr[(X_i - X_j)(Y_i - Y_j) > 0] - \Pr[(X_i - X_j)(Y_i - Y_j) < 0]
$$

$$
-1 \leq \tau \leq 1
$$

$$
\tau = 4 \int_0^1 \int_0^1 C(u, v)dC(u, v) - 1
$$

$$
\tau = \frac{1}{9} \delta - \frac{1}{900} \delta^3 + \frac{1}{52920} \delta^5 - \ldots \quad \text{Logistic model}
$$

$$
\tau = \frac{(\delta - 1)}{\delta} \quad \text{Frank’s copula}
$$
An overview of “Copulas”

These measures were introduced to generalize the “correlation coefficient” (e.g., Pearson’s product-moment correlation coefficient).

Problems of existence of “correlation coefficient”.
An overview of “Copulas”

**Empirical copula \( C_n \)**

Let \((x_k, y_k), k=1,\ldots,n\) be a sample of size \(n\) from a continuous bivariate distribution.

For \(i, j=1,\ldots, n\):

\[
C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{n_{ij}}{n}
\]

where \(n_{ij}\) is the number of sample pairs \((x, y)\) such that \(x \leq x_{(i)}\), \(y \leq y_{(i)}\), and \(x_{(i)}, y_{(j)}\) are the order statistics.

No assumptions on the margins of \(X, Y\) are needed to calculate \(C_n\).
An overview of "Copulas"
Modeling the statistical dependence of

- Storm rainfall intensity - Storm duration

\[ I - T \]
In the pioneer work on PRP, Eagleson (1972), assumed both the average intensity and the duration of the storm to be independent and exponentially distributed.

This assumption was preserved by Rodriguez-Iturbe et al. (1984); Rodriguez-Iturbe (1986); Bacchi et al. (1987) and by Klemeš (1978); Wood (1976), Wood and Hebson (1986); Cordova and Rodriguez-Iturbe (1983); Diaz-Granados et al. (1984); Raines and Valdes (1993).

Moreover, Bacchi et al (1987) showed how the PRP model with exponential marginals can provide in some cases a poor representation of the point rainfall process.
The possibility of representing storm duration by means of a long-tailed distribution was first explored by Rodriguez-Iturbe et al. (1987).

Robinson and Sivapalan (1997) modeled storm duration with a shifted exponential distribution and the conditional distribution of the average rainfall intensity with a gamma distribution.

Recently Menabde and Sivapalan (2000) modeled storm duration and average rainfall intensity by fat-tailed Levy distribution.
In literature, the intensity and duration are assumed, for simplicity, independent each. This hypothesis can not be realistic in some cases. Cordova and Rodriguez-Iturbe (1985) studied the effects of positive correlation between rainfall intensity and duration on storm surface runoff.

Bacchi et al. (1994) considered a bivariate exponential distribution introduced by Gumbel (1960) to model storm duration and average intensity.

Singh and Singh (1991) derived several bivariate distributions, with exponential margins, to represent the positive correlation between the variables.

Kurothe et al. (1997) used a bivariate distribution with exponential marginals to model also the negative correlation.
MENABDE AND SIVAPALAN: MODELING RAINFALL TIME SERIES AND EXTREMES
To model the statistical dependence (positive and negative) between storm rainfall intensity and storm duration, we consider here the Frank’s Copula.

\[
C(F_X, F_Y; \delta) = -\frac{1}{\delta} \ln \left\{ \frac{\eta - (1 - e^{-\delta F_X})(1 - e^{-\delta F_Y})}{\eta} \right\}
\]

\[
\eta = 1 - e^{-\delta}
\]
Marginal Distributions

Marginals of average rainfall intensity and storm duration

Generalized Pareto distribution

\[
F_I(i) = 1 - \left( 1 - \frac{k_I}{c_I} (i - b_I) \right)^{\frac{1}{k_I}}
\]

\[
F_T(t) = 1 - \left( 1 - \frac{k_T}{c_T} (t - b_T) \right)^{\frac{1}{k_T}}
\]

\[
f_I(i) = \frac{1}{c_I} \left( 1 - \frac{k_I}{c_I} (i - b_I) \right)^{\frac{1}{k_I} - 1}
\]

\[
f_T(t) = \frac{1}{c_T} \left( 1 - \frac{k_T}{c_T} (t - b_T) \right)^{\frac{1}{k_T} - 1}
\]

\[b_I, \text{ position parameter} \]
\[c_I, \text{ scale parameter} \]
\[k_I, \text{ shape parameter} \]

\[b_T, \text{ position parameter} \]
\[c_T, \text{ scale parameter} \]
\[k_T, \text{ shape parameter} \]
Model of storm rainfall

\[ F_{TI}(t,i) = -\frac{1}{\delta} \ln \left\{ \eta - \left( 1 - e^{-\delta \left( 1 - \frac{k_T}{c_T} (t - b_T) \right)^{\frac{1}{k_T}}} \right) \right\} / \eta \]

\[ f_{TI}(t,i) = \frac{\delta \eta e^{2 \delta + \delta \left( 1 - \frac{k_T}{c_T} (t - b_T) \right)^{\frac{1}{k_T}} + \delta \left( 1 - \frac{k_I}{c_I} (i - b_i) \right)^{\frac{1}{k_I}}} \left[ \eta - \left( 1 - e^{-\delta + \delta \left( 1 - \frac{k_T}{c_T} (t - b_T) \right)^{\frac{1}{k_T}}} \right) \right]^2 \cdot \left( 1 - \frac{k_T}{c_T} (t - b_T) \right)^{\frac{1}{k_T} - 1} \left( 1 - \frac{k_I}{c_I} (i - b_I) \right)^{\frac{1}{k_I} - 1} / c_T c_I \]
Parameters Estimation

Dependence parameter:

Measures of dependence $\tau$ or $\rho_S$

ML estimator

Parameters of marginal distributions:

L-moments, Hosking (1990)
Rainfall data from rain gauge located in Thyrrhenian Liguria

Hourly Data 1990-1997

Dry period = 7 hours
Threshold 1 mm/h

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_T$ [-]</td>
<td>-0.125</td>
</tr>
<tr>
<td>$c_T$ [hours]</td>
<td>11.459</td>
</tr>
<tr>
<td>$b_T$ [hours]</td>
<td>0</td>
</tr>
</tbody>
</table>

Intensity, mm/h

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_I$ [-]</td>
<td>-0.300</td>
</tr>
<tr>
<td>$c_I$ [mm/h]</td>
<td>1.534</td>
</tr>
<tr>
<td>$b_I$ [mm/h]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>( \delta )</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>( \tau ) (Kendall tau)</td>
<td>-0.092  -0.825</td>
</tr>
<tr>
<td>( \rho_s ) (Spearman rho)</td>
<td>-0.131  -0.787</td>
</tr>
<tr>
<td>ML</td>
<td>-         -0.857</td>
</tr>
</tbody>
</table>

\[ \text{C vs. } F_I, F_T \text{ and } C \]

\[ \text{Diagram showing data distribution in three dimensions.} \]
Control on rainfall volume

Montecarlo simulation

\[ V = T \cdot I \]
Monte Carlo simulation
I-D

The distribution of the rainfall volume, $P$

\[
F_P(p) = Pr[P \leq p] = \int\int_{\{i \leq p\}} f_{TI}(t,i)dtdi = \int_0^{+\infty} \left[ \int_0^{p/t} f_{TI}(t,i)di \right] dt
\]

\[
f_P(p) = \int_0^{+\infty} \frac{1}{t} f_{TI} \left( t, \frac{p}{t} \right) dt
\]

\[
f_P(p) \propto p^{k_I} \quad p \gg 1
\]
Coupling the 2-Copula with a probability distribution for the dry period
Two important aspects of rainfall storm modeling are considered

• The statistical dependence between duration and intensity is modeled using the concept of copula

• The storm duration and average rainfall intensity are modeled by Generalized Pareto distributions

• The model is tested using some rainfall data
Modeling the statistical dependence of

- Flood peak - Flood Volume
  \[ Q_{\text{max}} - V \]
To model the **positive** dependence between flood peak and flood volume, we consider here the Logistic Model

\[ C_\delta(u, v) = e^{-\left(\frac{(-\ln u)^\delta + (-\ln v)^\delta}{\delta}\right)} \]
Marginal Distributions

Marginals of flood peak and flood volume

Generalized Extreme Value distribution

\[
F_{Q_{\text{max}}} (q_{\text{max}}) = \exp \left[ - \left( 1 - \kappa_Q \frac{q_{\text{max}} - \varepsilon_Q}{\alpha_Q} \right)^{1/\kappa_Q} \right] \quad q > \varepsilon_{Q_{\text{max}}} + \frac{\alpha_{Q_{\text{max}}}}{\kappa_Q}
\]

\[
F_V (v) = \exp \left[ - \left( 1 - \kappa_V \frac{v - \varepsilon_V}{\alpha_V} \right)^{1/\kappa_V} \right] \quad v > \varepsilon_V + \frac{\alpha_V}{\kappa_V}
\]

\(\varepsilon\) position parameter, \(\alpha\) scale parameter, \(k\) shape parameter.
\[ F_{Q_{\text{max}}, V}(q, v) = P\left[ Q_{\text{max}} \leq q, V \leq v \right] = \]
\[ = \exp\left\{- \left[ \left( k_Q \frac{q - \varepsilon_Q}{\alpha_Q} \right)^{\delta/k_Q} + \left( k_V \frac{v - \varepsilon_V}{\alpha_V} \right)^{\delta/k_V} \right]^{1/\delta} \right\} \]
Case study: Ceppo Morelli dam

Anza catchment (a sub-basin of the Toce river basin, located in Northern Italy)

Area: 125 km²

Maximum water level: 782.5 m.s.l.m.

Dam crest level: 784 m.s.l.m.

Level of uncontrolled spillway: 780.75 m.s.l.m.

Length of uncontrolled spillway: 84 m.

Maximum water storage: 0.5·10⁶ m³

49 years of registration on dam
Case study
### Case study

<table>
<thead>
<tr>
<th>dependence</th>
<th>ML</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td>3.241</td>
<td>2.868</td>
</tr>
</tbody>
</table>
**Case study**

<table>
<thead>
<tr>
<th>V</th>
<th>ML</th>
<th>L-moments</th>
</tr>
</thead>
<tbody>
<tr>
<td>ε [$10^6 m^3$]</td>
<td>1.774</td>
<td>1.744</td>
</tr>
<tr>
<td>α [$10^6 m^3$]</td>
<td>1.544</td>
<td>1.620</td>
</tr>
<tr>
<td>k</td>
<td>-0.570</td>
<td>-0.564</td>
</tr>
</tbody>
</table>

### Q-V

<table>
<thead>
<tr>
<th>$Q_{max}$</th>
<th>ML</th>
<th>L-moments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon$ [m$^3$/s]</td>
<td>59.221</td>
<td>57.972</td>
</tr>
<tr>
<td>$\alpha$ [m$^3$/s]</td>
<td>35.676</td>
<td>33.689</td>
</tr>
<tr>
<td>k</td>
<td>-0.338</td>
<td>-0.420</td>
</tr>
</tbody>
</table>
From \((Q_{\text{max}}, V)\) to Flood Hydrograph

\[ q(t) = \begin{cases} 
1.335 \frac{Q_{\text{max}}^2}{V} t & 0 \leq t \leq T_p \\
1.6Q - 0.8 \frac{Q_{\text{max}}^2}{V} t & T_p \leq t \leq t_b 
\end{cases} \]

The Shape of Flood Hydrograph

Triangular hydrograph

Hydrograph with a Nash-IUH

\[ q(t) = \begin{cases} 
\frac{V}{t_0} \int_{t_0}^{t} \frac{1}{k\Gamma(n)} \left( \frac{t - \zeta}{k} \right)^{n-1} \ e^{-(t-\zeta)/k} \ d\zeta & 0 \leq t \leq t_0 \\
\frac{V}{t_0} \int_{t_0}^{t} \frac{1}{k\Gamma(n)} \left( \frac{t - \zeta}{k} \right)^{n-1} \ e^{-(t-\zeta)/k} \ d\zeta & t_0 \leq t 
\end{cases} \]
Adequacy of dam spillway and dam safety

1. Generation of initial reservoir level

2. Generation of couple \((Q_i, V_i)\) and the relative flood event

3. Direct Reservoir Routing

4. Distribution of highest water level during the flood
Conclusions

- A general model describing the possible bivariate behaviour of *flood-peak* and *flood-volume* is considered;

- A bivariate statistical procedure for the evaluation of *flood hydrograph* is presented;

- This procedure is used to check adequacy of dam *spillway* and thus the dam *safety*. 
Modeling the statistical dependence of

Flood peaks at different river sites
The regionalization methods of flood frequency analysis are generally based on the assumption of independence in space of the maximum annual flood peaks. However, the generation of floods is associated with precipitation events characterized by a mesoscale or synoptic extension, so flood peaks at neighboring sites are dependent in some degree. Because the amount of information from \( N \) dependent series is less than \( N \) times that from a single series, the available information in a region can be seriously affected by spatial dependence. This can introduce further uncertainties in flood frequency regionalization. The present paper deals with a multivariate probability model with generalized extreme value margins. The model can account for the presence of intersite dependence in the flood frequency regionalization exercise. Examples of application to northwestern Italy are presented and discussed.
To model the positive statistical dependence among flood peak at different river sites, we consider a simple Logistic Model

\[ C_\delta(u_1, u_2, u_3, ..., u_N) = e^{-\left[-\left(\sum_{i=1}^{N} \ln u_i \right)\delta \right]^{1/\delta}} \]

Simple N-variate model. It assumes that the dependence between each couple of sites is represented through the \( \delta \) parameter. It can be viewed as a measure of the regional dependence.
It is applied to the normalized maximum annual flood peak, \( X_i \), defined as \( X_i = \frac{Q_i}{\mathbb{E}[Q_i]} \) the ratio between the maximum annual flood peak at a river site \( i \), \( Q_i \), and its expected value, \( \mathbb{E}[Q_i] \).

The generalized extreme value distribution, GEV, is considered as marginal of the normalized variable \( X_i \) with \( i=1, \ldots, N \). Its cumulative distribution function, cdf, is

\[
F_{X_i}(x_i) = \exp\left\{ - \left[ 1 - k'(x_i - \epsilon')/\alpha' \right]^{1/k'} \right\} \quad \text{with } i=1, \ldots, N
\]

where the marginal parameters \( \epsilon' \), \( \alpha' \) and \( k' \), respectively, position, scale and shape parameter, are the same for each site within the homogeneous region.
Multivariate Extreme Value model for flood frequency regionalization in presence of spatial dependence

\[ F_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = e^{-\left\{ \sum_{i=1}^{N} \left[ 1 - k'(x_i - \varepsilon')/\alpha' \right]^{\delta'/k'} \right\}^{1/\delta}} \]

\[ f_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = \frac{\partial^N F_{X_1,\ldots,X_N}}{\partial x_1 \cdots \partial x_N} = \prod_{i=1}^{N} \left( \frac{\nu_i^{\delta-1}}{\alpha'} \right) \cdot \frac{z^{1-N\delta}}{\exp(z)} \cdot W_N(z) \]

where

\[ \nu_i = \left[ 1 - k'(x_i - \varepsilon')/\alpha' \right]^{1/k'} \]

\[ z = \left( \sum_{i=1}^{N} \nu_i^\delta \right)^{1/\delta} \]

\[ W_N(z) = [(N-1)\delta + z - 1] \cdot W_{N-1}(z) - z \frac{\partial W_{N-1}(z)}{\partial z} \]
Model is characterized by 4 parameters: $\varepsilon'$, $\alpha'$, $k'$ and $\delta$

These are estimated using ML method

$$ll = \ln f_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = -N \ln \alpha' + (\delta - k') \cdot \sum_{i=1}^{N} \ln y_i + (1 - N\delta) \cdot \ln z - z + \ln W_N(z)$$

Initial estimates of the marginal parameters, $\varepsilon'$, $\alpha'$, $k'$, can be obtained applying the probability weighted moments (PWM) technique [Hosking et al., 1985] on the regional normalized sample.

An initial estimate of the dependence parameter, $\delta$, can be evaluated through an estimate of the Kendall's $\tau$ measure of dependence, Kendall [1937].
Case study
Case study

Region A - structure of dependence
## Case study

### Estimation of model parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Region A (N=14)</th>
<th>Region B (N=14)</th>
<th>Region C (N=27)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>multivariate</td>
<td>univariate</td>
<td>multivariate</td>
</tr>
<tr>
<td>(\varepsilon')</td>
<td>0.760</td>
<td>0.740</td>
<td>0.656</td>
</tr>
<tr>
<td>(\alpha')</td>
<td>0.400</td>
<td>0.371</td>
<td>0.395</td>
</tr>
<tr>
<td>(k')</td>
<td>-0.223</td>
<td>-0.104</td>
<td>-0.336</td>
</tr>
<tr>
<td>(\delta)</td>
<td>1.412</td>
<td>1.347</td>
<td>1.350</td>
</tr>
</tbody>
</table>
### Case study

<table>
<thead>
<tr>
<th>Region</th>
<th>T</th>
<th>$\hat{x}_T$</th>
<th>$\text{var}[\hat{x}_T] \times (10^{-3})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>multivariate</td>
<td>univariate</td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>1.472</td>
<td>1.343</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.929</td>
<td>1.681</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2.445</td>
<td>2.032</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>3.248</td>
<td>2.527</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>3.970</td>
<td>2.931</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>4.809</td>
<td>3.364</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>1.317</td>
<td>1.425</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.800</td>
<td>1.982</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2.383</td>
<td>2.666</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>3.363</td>
<td>3.835</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>4.311</td>
<td>4.985</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>5.486</td>
<td>6.433</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>1.334</td>
<td>1.406</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.813</td>
<td>1.953</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2.377</td>
<td>2.614</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>3.301</td>
<td>3.728</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>4.171</td>
<td>4.806</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>5.226</td>
<td>6.145</td>
</tr>
</tbody>
</table>

- **De Michele C.,** G. Salvadori, M. Canossi, A. Petaccia, and R. Rosso, Bivariate statistical approach to spillway design flood, *accepted for publication, ASCE J. Hydrol. Eng.*