MULTIVARIATE RETURN PERIODS IN EARTH SCIENCES: A COPULA APPROACH

Gianfausto SALVADORI

Dipartimento di Matematica e Fisica
Università del Salento
Lecce (Italy)

gianfausto.salvadori@unisalento.it

http://www.unisalento.it/people/gianfausto.salvadori

Columbia University Statistics Department
New York, October 11–12, 2013

http://dependence2013.wikischolars.columbia.edu/Copula+Conference

MULTIVARIATE RETURN PERIOD

Introduction

The concept of RETURN PERIOD (RP) is adopted in Earth Sciences (e.g., hydrology, coastal & off-shore engineering, volcanology, seismology, ...) as a common criterion for design purposes and sizing works, and may also provide a means for risk analysis and assessment.

Definition (Informal)

The traditional definition of the RP is as "the average time elapsing between two successive occurrences of a prescribed (dangerous) event".

Basic Assumptions

1. The region of dangerous events depends upon the RP.
2. The structure of interest is sized using the notion of RP.

★ The RP is analogous to the "Value-at-Risk (VaR)" used in Finance.

THE MULTIVARIATE RP FRAMEWORK

The ingredients

- We consider a sequence $\mathscr{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$ of i.i.d. d-dimensional random vectors $\mathbf{X}_i$'s, $d \geq 1$, describing the phenomenon.
- Each vector $\mathbf{X}_i$ has a multivariate distribution $\mathbf{F} = \mathbf{C}(F_1, \ldots, F_d)$ (Sklar's Theorem), with suitable margins F_i's and d-copula $\mathbf{C}$.
- The "Failure Region" is a set $\mathscr{D} \subset \mathbf{R}^d$ collecting all the events judged to be dangerous according to a suitable criterion (RP).

Definition (μ: timing of the phenomenon)

The constant $\mu > 0$ will denote the average inter-arrival time of the occurrences in $\mathscr{X}$, viz. the average time elapsing between successive occurrences $\mathbf{X}_i$ and $\mathbf{X}_{i+1}$. For instance, $\mu = 1$ year for annual maxima.

THE MULTIVARIATE RP FRAMEWORK

General RP: for an overview see **[Gräler et al., HESS, 2013]**

Definition (General RP [Salvadori & De Michele, WRR, 2004])

The Return Period $T_{\mathscr{D}}$ associated with the dangerous set $\mathscr{D}$ is given by

$$T_{\mathscr{D}} = \mu / \mathbb{P}\{\mathbf{X} \in \mathscr{D}\}. \quad (1)$$

The above Definition is a very general one: most of the approaches already available in literature are particular cases of the present one.

Example (Traditional univariate approach)

1. The RP T is fixed *a priori* by prescribed regulation constraints.
2. The Failure Region is $\mathscr{D} = \{X > x_T\}$, which depends upon T.
3. The probability of $\mathscr{D}$ is $\mathbb{P}\{X \in \mathscr{D}\} = 1 - F_X(x_T) = \mu / T$.
4. The design quantile x_T is calculated as $x_T = F_X^{-1}(1 - \mu / T)$.

HYDROLOGICAL EXAMPLE

Univariate case (Droughts & Floods)

Let X be a univariate r.v. modeling the river discharge, with distribution F_X, and let x^* denote a prescribed critical threshold value. Then,

1. if DROUGHTS are of concern, x^* may represent a small value of discharge, and the critical occurrences are those for which $X < x^*$ (viz., $\mathscr{D}^{\text{dr.}} = [0, x^*)$, with probability $p_{\mathscr{D}}^{\text{dr.}} = F_X(x^*)$);
2. if FLOODS are of concern, x^* may represent a large value of discharge, and the critical occurrences are those for which $X > x^*$ (viz., $\mathscr{D}^{\text{fl.}} = (x^*, \infty)$, with probability $p_{\mathscr{D}}^{\text{fl.}} = 1 - F_X(x^*)$).

Since $T_{\mathscr{D}} = \mu / p_{\mathscr{D}}$, the corresponding univariate RP's are

1. $T_{\mathscr{D}}^{\text{dr.}} = \mu / p_{\mathscr{D}}^{\text{dr.}} = \mu / F_X(x^*)$ in the droughts' case;
2. $T_{\mathscr{D}}^{\text{fl.}} = \mu / p_{\mathscr{D}}^{\text{fl.}} = \mu / (1 - F_X(x^*))$ in the floods' case.

- Case 2 corresponds to the "traditional" definition of (univariate) RP.

THE MULTIVARIATE RP FRAMEWORK

Consistent extension to the multivariate case **[Salvadori et al., HESS, 2011]**

The notion of critical layer $\mathscr{L}_t^{\mathbf{F}}$ provides a multivariate threshold.

Definition (Critical layer)

Given a d-dimensional distribution $\mathbf{F} = \mathbf{C}(F_1, \ldots, F_d)$ and a critical level $t \in (0,1)$, the critical layer $\mathscr{L}_t^{\mathbf{F}}$ of level t is defined as

$$\mathscr{L}_t^{\mathbf{F}} = \{\mathbf{x} \in \mathbf{R}^d : \mathbf{F}(\mathbf{x}) = t\}. \tag{2}$$

Note

- $\mathscr{L}_t^{\mathbf{F}}$ is the iso-hyper-surface (with dimension $d-1$) where $\mathbf{F} = t$: for instance, $\mathscr{L}_t^{\mathbf{F}}$ is a (iso)line for bivariate distributions, a (iso)surface for trivariate ones, and so on.
- For any $\mathbf{x} \in \mathbf{R}^d$, the critical layer $\mathscr{L}_{t=\mathbf{F}(\mathbf{x})}^{\mathbf{F}}$ supporting $\mathbf{x}$ is *unique*.

THE MULTIVARIATE RP FRAMEWORK

Events' space partition **[Salvadori et al., HESS, 2011]**

As for the univariate case, the critical layer $\mathscr{L}_t^{\mathbf{F}}$ is a multivariate threshold, and partitions $\mathbf{R}^d$ into three non-overlapping exhaustive regions.

1. The sub-critical region $R_t^{<} = \{\mathbf{x} : \mathbf{F}(\mathbf{x}) < t\}$;
2. The critical layer $\mathscr{L}_t^{\mathbf{F}}$, where $\mathbf{F} \equiv t$;
3. The super-critical region $R_t^{>} = \{\mathbf{x} : \mathbf{F}(\mathbf{x}) > t\}$.

Note

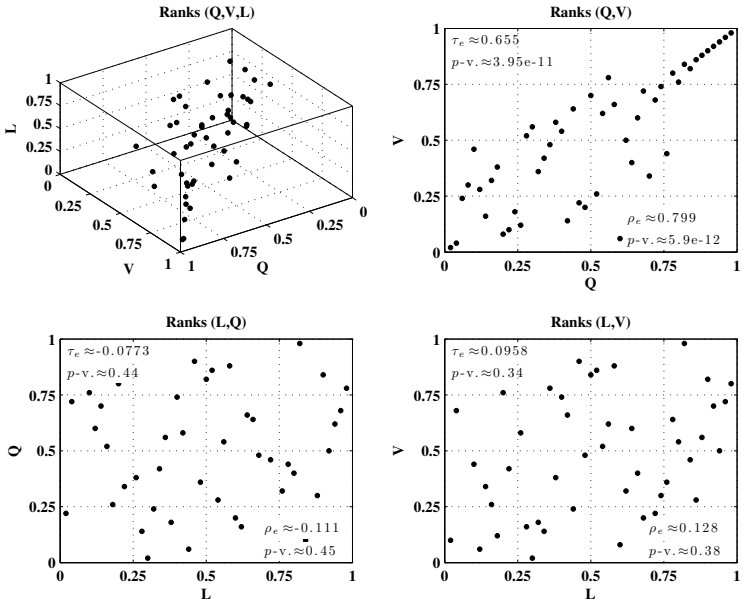
At any realization of the phenomenon, the random occurrence $\mathbf{x}$ lies

1. either in the sub-critical region $R_t^{<}$,
2. or on the critical layer $\mathscr{L}_t^{\mathbf{F}}$,
3. or in the super-critical region $R_t^{>}$.

EXAMPLE

Ceppo Morelli dam, Italy **[Salvadori et al., HESS, 2011]**

DATA



MAIN VARIABLES

- max annual flood peak Q
- max annual flood volume V
- initial reservoir water level L

- $Q \sim F_Q$: GEV; $V \sim F_V$: GEV
- (Q, V) are fairly concordant
- L is independent of (Q, V)
- $(Q, V) \sim \mathbf{C}_{QV}$: Gumbel EVC

The following trivariate copula is used to model the triple (Q, V, L):

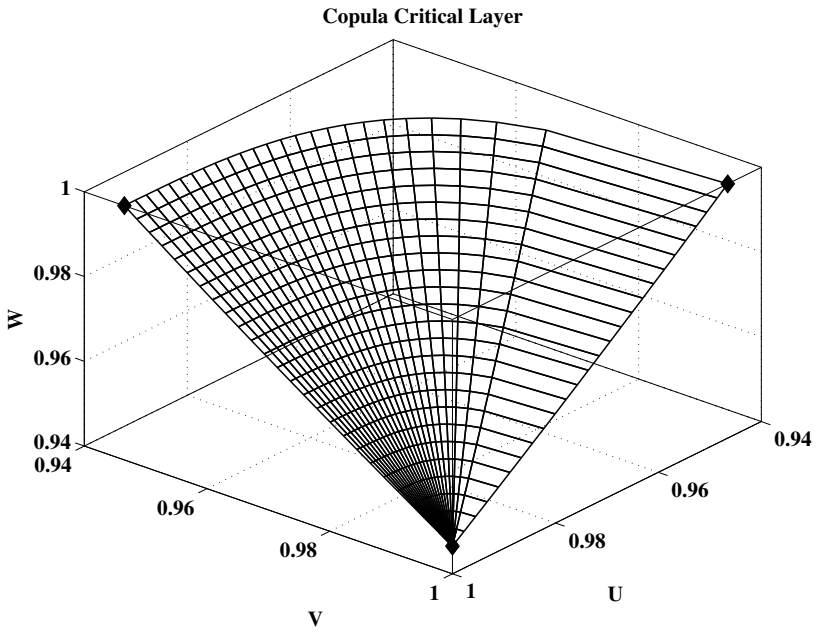
$$\mathbf{C}_{QVL}(u, v, w) = \mathbf{C}_{QV}(u, v) \cdot w.$$

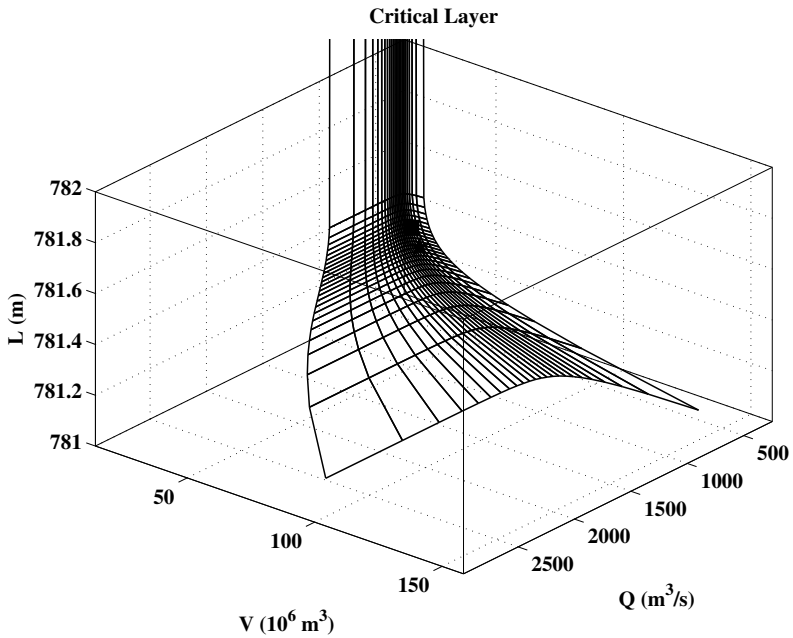
(1) $\mathbf{F}_{QV} = \mathbf{C}_{QV}(F_Q, F_V)$ is MEV. (2) $\mathbf{C}_{QVL}$ is EVC (it is max-stable). (3) However, since F_L is not GEV, $\mathbf{F}_{QVL} = \mathbf{C}_{QVL}(F_Q, F_V, F_L)$ is not MEV.

EXAMPLE

Ceppo Morelli dam, Italy **[Salvadori et al., HESS, 2011]**

2-D Critical layers $\mathscr{L}_{t^*}^{\mathbf{C}_{QVL}}$ and $\mathscr{L}_{t^*}^{\mathbf{F}_{QVL}}$, with level $t^* \approx 0.946519$, sub-critical region $R_{t^*}^{<}$, and super-critical region $R_{t^*}^{>}$.





On the 2-D critical layers, $\mathbf{C}_{QVL} \equiv t^*$ and $\mathbf{F}_{QVL} \equiv t^*$. Via the Probability Integral Transform, the correspondence $\mathscr{L}_{t^*}^{\mathbf{C}_{QVL}} \leftrightarrow \mathscr{L}_{t^*}^{\mathbf{F}_{QVL}}$ is one-to-one.

THE MULTIVARIATE RP FRAMEWORK

The Kendall's Measure $K_{\mathbf{C}}$ [Genest & Rivest, JASA-SPL, 1993-2001]

It is then essential to provide a probability measure for $R_t^<$ (or $R_t^>$). The Kendall's measure represents a fundamental tool for introducing a suitable COPULA-BASED definition of RP for multivariate events.

Definition (Kendall's Measure)

Let $\mathbb{I} = [0,1]$. The Kendall's measure $K_{\mathbf{C}} : \mathbb{I} \to \mathbb{I}$ is given by

$$K_{\mathbf{C}}(t) = \mathbb{P}\{\mathbf{F}(X_1,\ldots,X_d) \leq t\} = \mathbb{P}\{\mathbf{C}(F_1(X_1),\ldots,F_d(X_d)) \leq t\}, \quad (3)$$

where $t \in \mathbb{I}$, and the X_i's have joint law $\mathbf{F} = \mathbf{C}(F_1,\ldots,F_d)$.

Note (Practical interpretation)

$K_{\mathbf{C}}(t)$ is the probability that a multivariate realization simulated from $\mathbf{F}$ appears in the sub-critical region $R_t^<$ (multivariate quantile relationship).

THE MULTIVARIATE RP FRAMEWORK

(super-critical) Kendall's RP **[Salvadori et al., HESS, 2011]**

The following notion of multivariate RP is analogous to the 1-D one.

Definition (Kendall's RP)

- Let $\mathbf{X}$ be a multivariate r.v. with distribution $\mathbf{F} = \mathbf{C}(F_1, \ldots, F_d)$.
- Let $\mathscr{L}_t^{\mathbf{F}}$ be the critical layer supporting $\mathbf{X}$, with $t = \mathbf{F}(\mathbf{X})$.
- Let $R_t^{>}$ be the corresponding super-critical region.

Then, the (super-critical) Kendall's RP (KRP) $\kappa_{\mathbf{X}}$ associated with $\mathbf{X}$ is

$$\kappa_{\mathbf{X}} = \frac{\mu}{\mathbb{P}\{\mathbf{X} \in \mathscr{D}\}} = \frac{\mu}{\mathbb{P}\left\{\mathbf{X} \in R_t^{>}\right\}} = \frac{\mu}{1 - K_{\mathbf{C}}(t)}. \tag{4}$$

Note

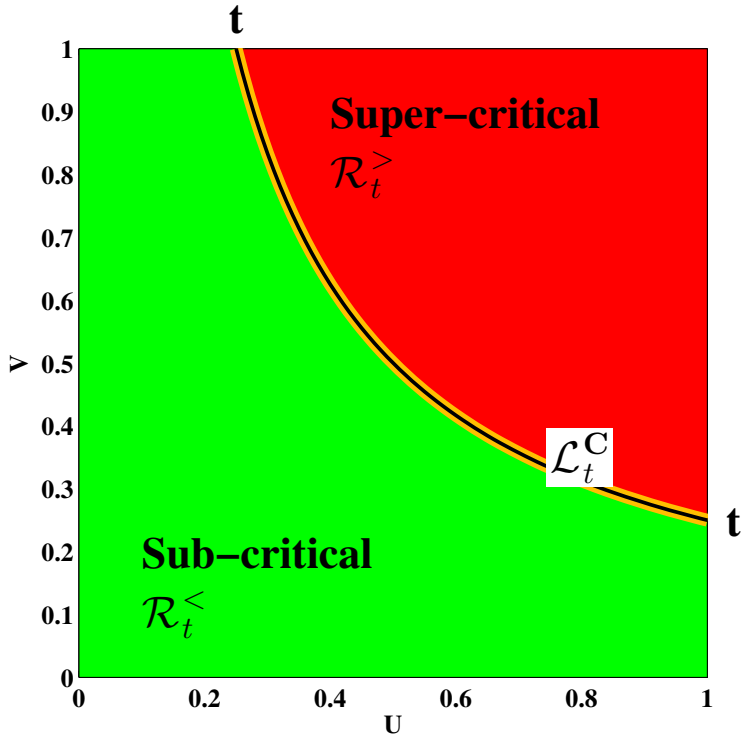
All the occurrences lying on the critical layer $\mathscr{L}_t^{\mathbf{F}}$, with $t = \mathbf{F}(\mathbf{X})$, share the same KRP $\kappa_{\mathbf{X}}$: thus, there are ∞^{d-1} occurrences with KRP $= \kappa_{\mathbf{X}}$.

THE MULTIVARIATE RP FRAMEWORK

(super-critical) Kendall's RP **[Salvadori et al., HESS, 2011]**

Thanks to the Probability Integral Transform, the notions of sub-critical region, critical layer, and super-critical region are meaningful also in the Copula domain, where $\mathbf{F} = \mathbf{C}(F_1, \ldots, F_d)$.

PARTITIONING



1-D analogies

- $\kappa_{\mathbf{X}}$ is a function of the critical probability level $t = \mathbf{F}(\mathbf{X})$
- All the "safe" occurrences lying in the sub-critical region have a KRP $< \kappa_{\mathbf{X}}$
- All the "alert" occurrences lying on $\mathscr{L}_t^{\mathbf{C}}$ have a KRP $= \kappa_{\mathbf{X}}$
- All the "dangerous" occurrences lying in the super-critical region have a KRP $> \kappa_{\mathbf{X}}$

THE MULTIVARIATE RP FRAMEWORK

Multivariate Kendall's quantile **[Salvadori et al., HESS, 2011]**

A consistent notion of multivariate quantile can be given as follows.

Definition (Multivariate Kendall's quantile)

Given a d-dimensional distribution $\mathbf{F} = \mathbf{C}(F_1, \ldots, F_d)$, and a critical probability level $p \in \mathbb{I}$, the multivariate Kendall's quantile $t \in \mathbb{I}$ of order p is

$$t = \inf\{x \in \mathbb{I}\colon K_{\mathbf{C}}(x) = p\} = K_{\mathbf{C}}^{(-1)}(p). \tag{5}$$

- Given a "regulation" RP T, the critical probability level is calculated as $p(T) = 1 - \mu/T$, and the corresponding critical level is $t(T) = K_{\mathbf{C}}^{(-1)}(p(T))$, which identifies the critical layer $\mathscr{L}^{\mathbf{F}}_{t=t(T)}$.
- In a multivariate simulation of size N extracted from $\mathbf{F}$, Np realizations are expected to lie in $R^{<}_{t=t(T)}$, and the others in $R^{>}_{t=t(T)}$.

THE MULTIVARIATE RP FRAMEWORK

Practical multivariate procedure (analogous to the univariate one)

1. A critical design RP, say T, is fixed by regulation constraints (e.g., 5/10 y for sewers, 50/100 y for harbors, 1000 y for dams).
2. In order to identify the super-critical region of dangerous events $\mathscr{D} \equiv R_t^{>}$, first its probability $p_{\mathscr{D}}$ is calculated as

$$T = \mu / \mathbb{P}\{\mathbf{X} \in \mathscr{D}\} = \mu / p_{\mathscr{D}} \quad \Rightarrow \quad p_{\mathscr{D}} = \mu / T .$$

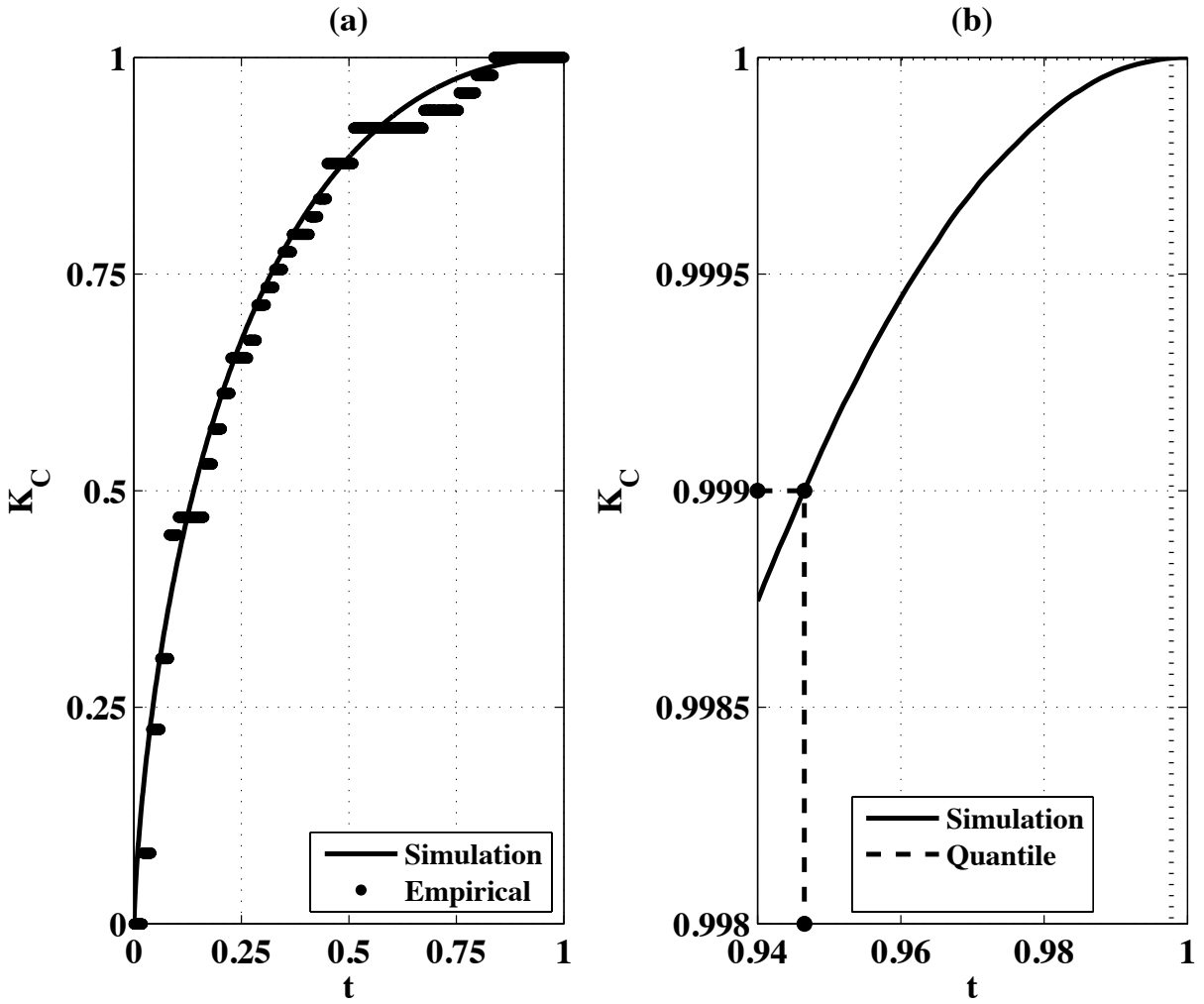
3. Then, the level t^* of the corresponding critical layer $\mathscr{L}_{t^*}^{\mathbf{F}}$ (i.e., the critical Kendall's quantile t^*) is calculated by inverting $K_{\mathbf{C}}$:

$$1 - K_{\mathbf{C}}(t^*) = p_{\mathscr{D}} \quad \Rightarrow \quad t^* = K_{\mathbf{C}}^{(-1)}(1 - \mu / T) .$$

In general, the calculation of t^* (i.e., the inversion of $K_{\mathbf{C}}$) can be carried out via Monte Carlo techniques **[Salvadori et al., HESS, 2011]**.

EXAMPLE

Ceppo Morelli dam, Italy **[Salvadori et al., HESS, 2011]**



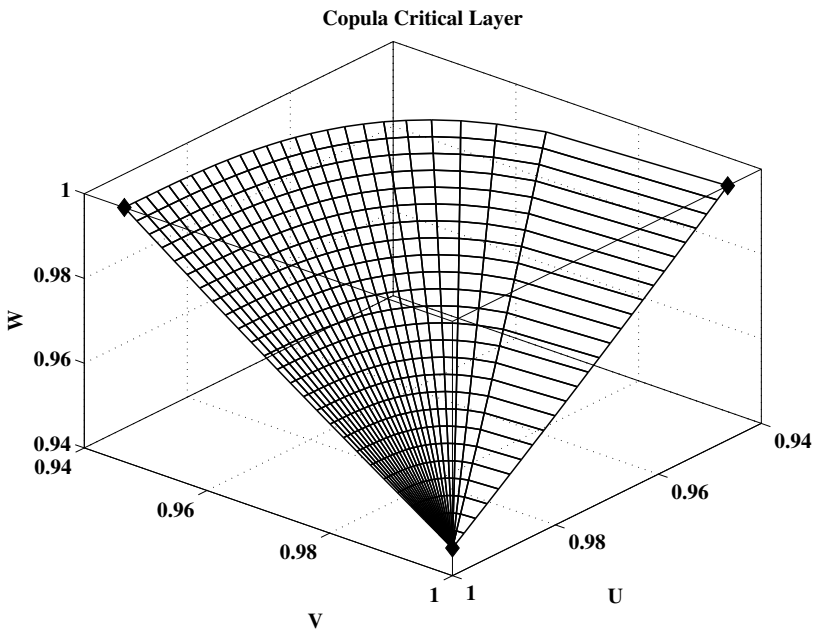
(*a*) Simulation-based estimate of $K_{\mathbf{C}}$ associated with the 3-copula $\mathbf{C}_{QVL}$, and empirical estimate of $K_{\mathbf{C}}$.

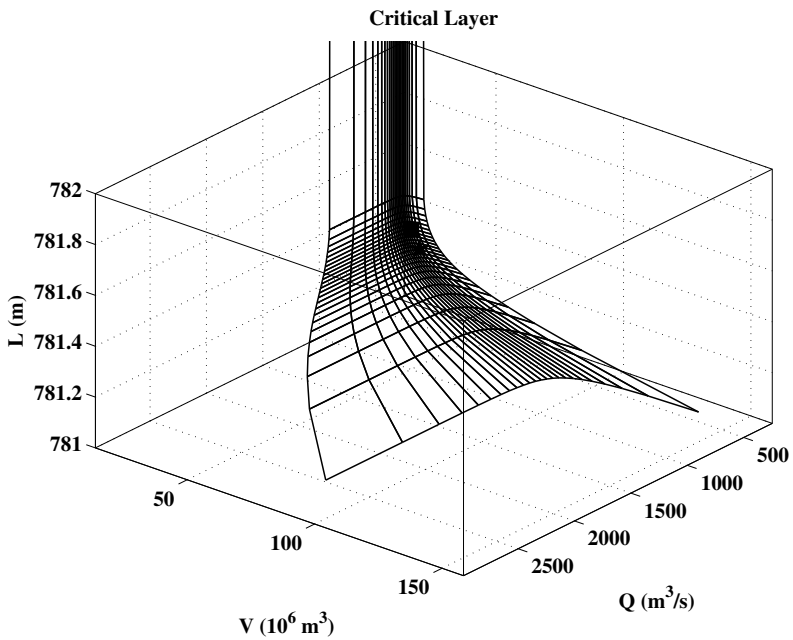
(*b*) Estimate of the multivariate quantile $t^* \approx 0.946519$ associated with $p = 0.999$, corresponding to a Millenary KRP: the super-critical region has probability 0.001.

EXAMPLE

Ceppo Morelli dam, Italy **[Salvadori et al., HESS, 2011]**

Critical layers $\mathscr{L}_{t^*}^{\mathbf{C}_{QVL}}$ and $\mathscr{L}_{t^*}^{\mathbf{F}_{QVL}}$, with Millenary level $t^* \approx 0.946519$.



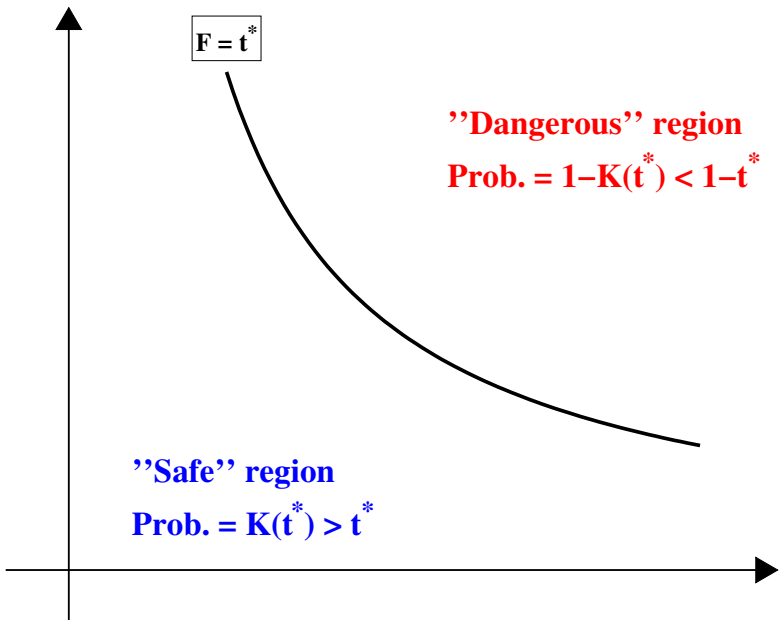


All the occurrences lying on the critical layers $\mathscr{L}_{t^*}^{\mathbf{C}_{QVL}}$ and $\mathscr{L}_{t^*}^{\mathbf{F}_{QVL}}$ have a 1000-years KRP: then, Millenary design triples can be calculated...

THE MULTIVARIATE RP FRAMEWORK

The Multivariate RP's are ruled by $K_{\mathbf{C}}$

Kendall's approach



Events' space partition

Common FALLACY

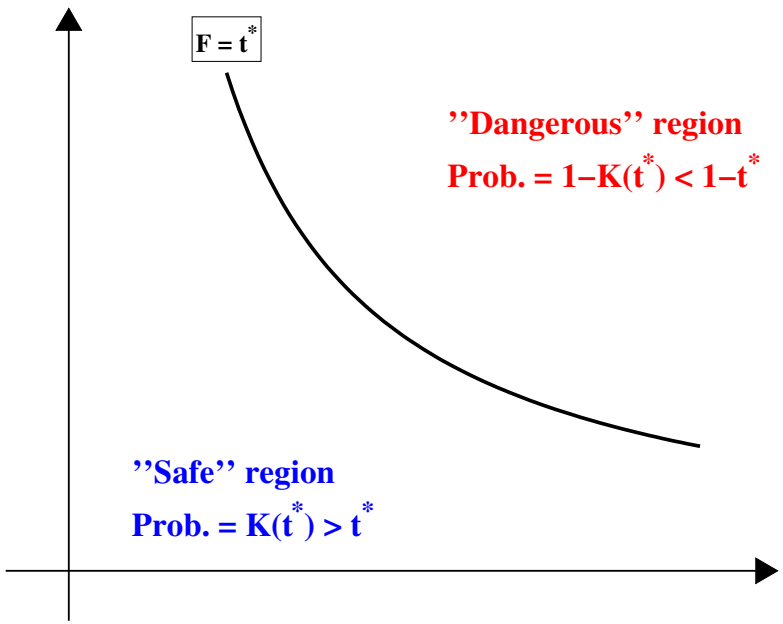
It is a common error to confuse the value of **C** (or **F**) with the probabilities induced by **C** (or **F**) on the relevant regions:

- on the critical layer $\mathscr{L}^{\mathbf{C}}_{t^*}$ (or $\mathscr{L}^{\mathbf{F}}_{t^*}$) it is $\mathbf{C} = \mathbf{F} = t^*$, but
- the sub-critical region has probability $K_{\mathbf{C}}(t^*) > t^*$, and
- the super-critical region has probability $1 - K_{\mathbf{C}}(t^*) < 1 - t^*$.

THE MULTIVARIATE RP FRAMEWORK

The SURVIVAL improvement **[Salvadori et al., WRR, 2013]**

KRP approach



"Safe"/"Dangerous" regions

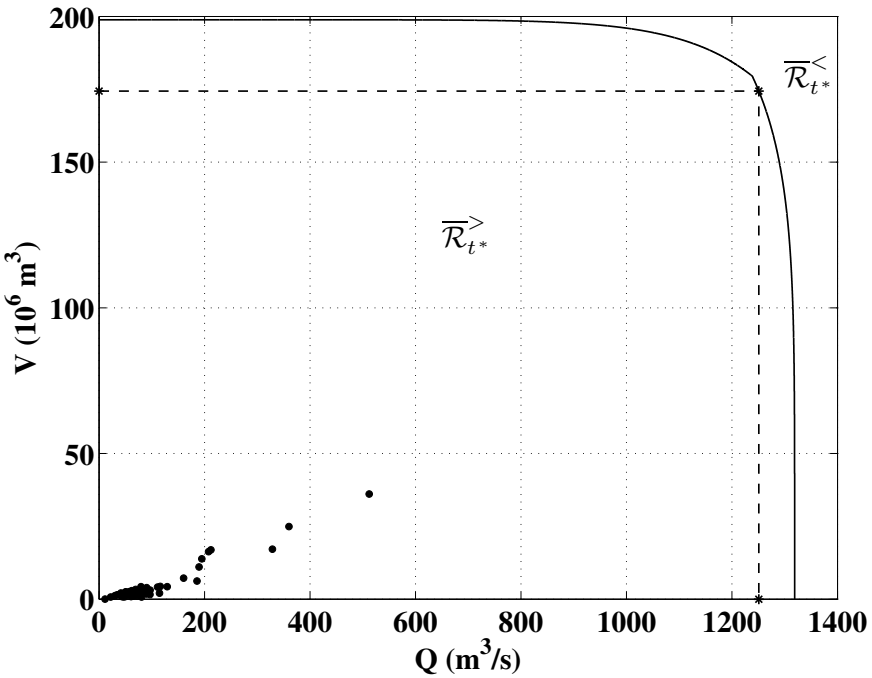
KRP drawbacks

- The "safe" region may be unbounded, i.e. the variables may take very large values, although with negligible probabilities.
- This fact may be counter-intuitive, when dealing with "safe" occurrences.

THE MULTIVARIATE RP FRAMEWORK

The SURVIVAL improvement **[Salvadori et al., WRR, 2013]**

Survival-KRP approach



Bounded "Safe" region

S-KRP advantages

- A possible solution is to use the survival function $\overline{\mathbf{F}}$ instead of the distribution **F**.
- The critical layers (i.e. the multivariate thresholds) are concave instead of convex.
- The "safe" region turns out to be bounded: the S-KRP may provide feasible solutions.

THE MULTIVARIATE RP FRAMEWORK

The DYNAMIC Return Period (DRP) **[De Michele et al., WRR, 2013]**

The notion of multivariate RP could be used as a tool for multivariate real-time risk assessment via the DYNAMIC Return Period (DRP).

Example (Multivariate real-time drought assessment)

Droughts are characterized by several non-independent variables: the Duration D, the Intensity I, the Severity S, and so on.

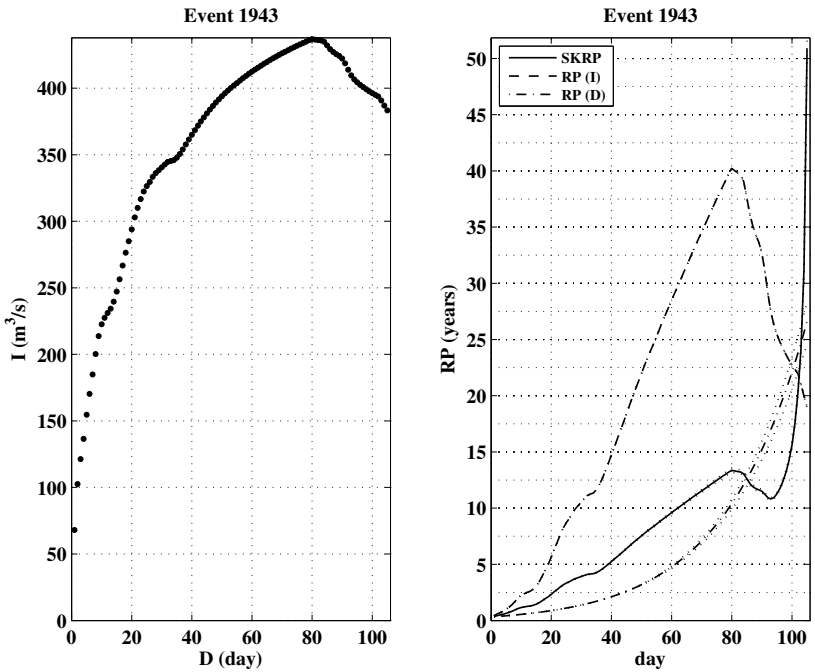
In the traditional univariate approach, "Alert", "Alarm", and "Emergency" decreasing discharge thresholds (m^3/s) are decided by the Watershed Authority on a physical base: the Intensity I is the "regulation" variable.

Instead, the multivariate approach suggests to check the temporal evolution of the (S)KRP of the pair (I, D) — called Dynamic Return Period: this may provide a better assessment of the risk, and a real-time tool to decide if, and when, mitigation interventions should be carried out.

THE MULTIVARIATE RP FRAMEWORK

The DYNAMIC Return Period (DRP) **[De Michele et al., WRR, 2013]**

Drought Episode 1943



Po's at Pontelagoscuro (Italy)

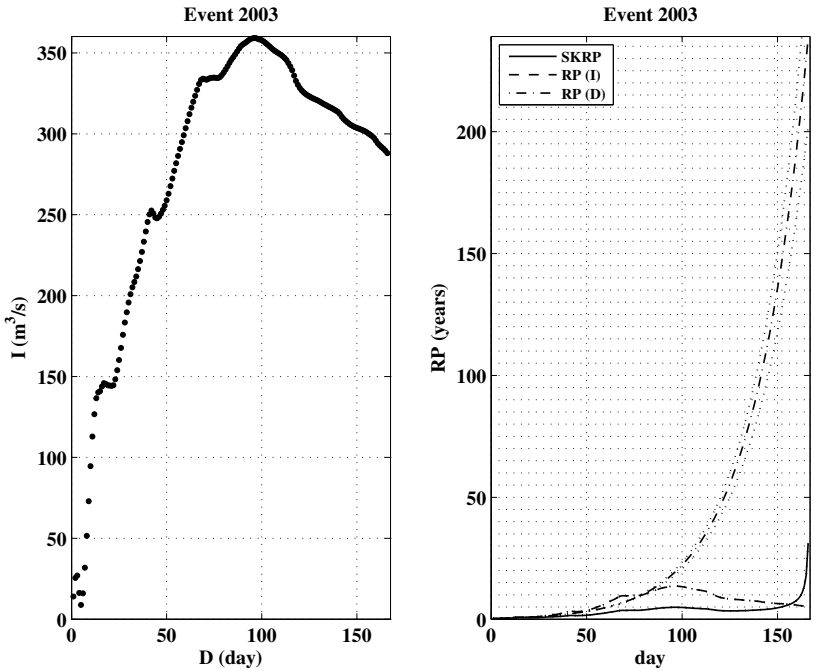
DRP analysis (no extremes)

- Neither I_k nor D are really too extreme, with maximum RP's smaller than 40 and 25 years, respectively.
- By the end of the drought, I_k decreases, but the DRP explodes ($>$ 50 years), due to the combined action of D.
- This latter information cannot be conveyed by the analysis of the univariate marginal RP's of I_k and D only.

THE MULTIVARIATE RP FRAMEWORK

The DYNAMIC Return Period (DRP) **[De Michele et al., WRR, 2013]**

Drought Episode 2003



Po's at Pontelagoscuro (Italy)

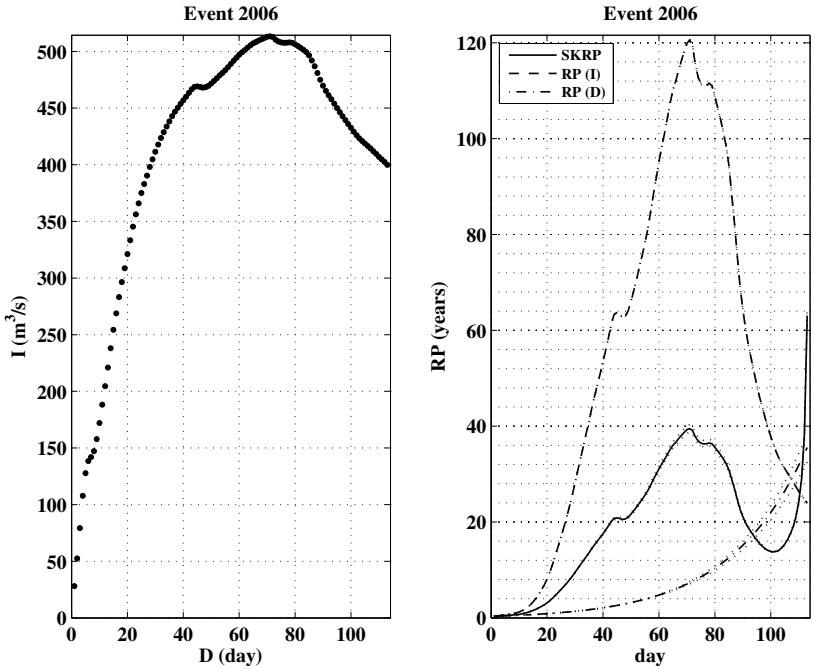
DRP analysis (*D* extreme)

- I_k is moderate (max RP of 15 years), whereas *D* is extreme (more than 5 months, max RP of 250 years).
- The DRP is always small (about 5 years), and increases only by the end of the drought, indicating that some further intervention is required.
- The DRP provides a tool to decide when small intensities become dangerous because of too a prolonged duration.

THE MULTIVARIATE RP FRAMEWORK

The DYNAMIC Return Period (DRP) **[De Michele et al., WRR, 2013]**

Drought Episode 2006



Po's at Pontelagoscuro (Italy)

DRP analysis (I_k extreme)

- D is not extreme (max RP of 40 years), whereas I_k is extreme, with centenary RP's.
- The DRP increases rapidly, attaining the level 40 years in only ten weeks, and it stays large, with a further abrupt raise around the 15-th week.
- The DRP indicates that the intervention has not been successful, and should have been anticipated in the early weeks.

THE MULTIVARIATE DESIGN FRAMEWORK

Introduction

The multivariate approach generally fails to provide a natural solution to the problem of identifying a unique design occurrence on the critical layer $\mathscr{L}_t^{\mathbf{F}}$ with given (S)KRP. The idea is to introduce a suitable function $\mathbf{w}$ that "weighs" the occurrences lying on $\mathscr{L}_t^{\mathbf{F}}$.

Definition (Design realization)

Let $\mathbf{w}\colon \mathscr{L}_t^{\mathbf{F}} \to [0,\infty)$ be a weight function defined on the critical layer $\mathscr{L}_t^{\mathbf{F}}$. The design realization $\delta_{\mathbf{w}} \in \mathscr{L}_t^{\mathbf{F}}$ is defined as

$$\delta_{\mathbf{w}}(t) = \operatorname*{argmax}_{\mathbf{x} \in \mathscr{L}_t^{\mathbf{F}}} \mathbf{w}(\mathbf{x}), \tag{6}$$

provided that the argmax exists and is finite.

The practitioner can then freely choose the selection criterion (i.e., the weight function $\mathbf{w}$) that best fits the practical needs.

THE MULTIVARIATE DESIGN FRAMEWORK

Most-Likely design realization

A natural approach consists in taking into account the density **f** of **F** over $\mathscr{L}_t^{\mathbf{F}}$ (if it is well defined). The restriction of **f** is not a proper density, but it induces a (weak) form of likelihood over $\mathscr{L}_t^{\mathbf{F}}$, and can be used to spot the occurrences on $\mathscr{L}_t^{\mathbf{F}}$ that are (relatively) "more likely" than others.

Definition (Most-Likely weigh function)

The Most-Likely weight function $\mathbf{w}_{\mathrm{ML}}$ is defined as

$$\mathbf{w}_{\mathrm{ML}}(\mathbf{x}) = \mathbf{f}(\mathbf{x}), \tag{7}$$

where **f** is the density of $\mathbf{F} = \mathbf{C}(F_1, \ldots, F_d)$.

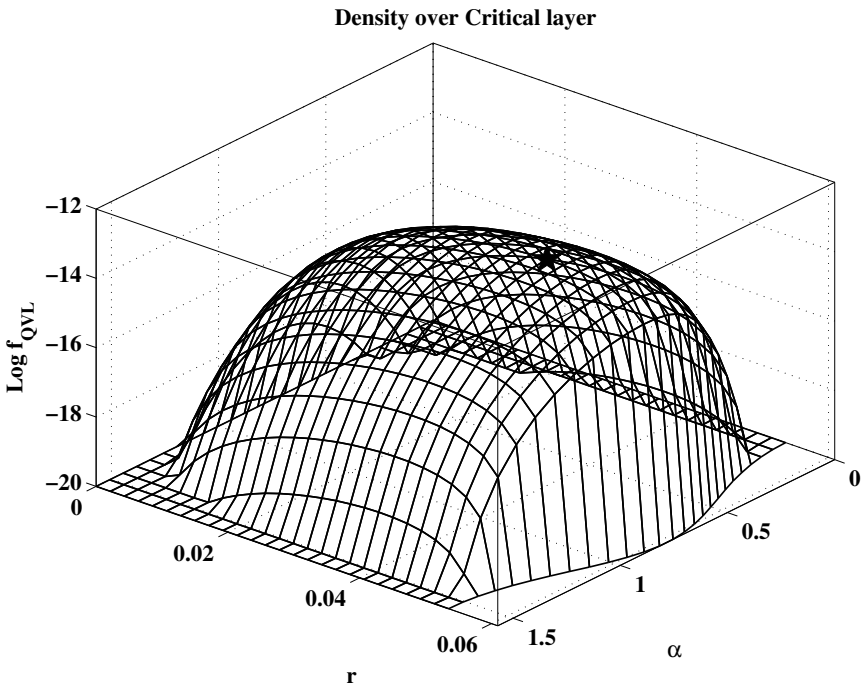
Note

Using the copula density **c**: $\mathbf{f}(\mathbf{x}) = \mathbf{c}(F_1(x_1), \ldots, F_d(x_d)) \cdot \prod_{i=1}^{d} f_i(x_i)$.

EXAMPLE

Ceppo Morelli dam, Italy **[Salvadori et al., HESS, 2011]**

Millenary critical layer



Most-Likely design realization

- The copula $\mathbf{C}_{QVL}$ is trivariate.
- The millenary critical layer is 2D, and the density $\mathbf{f}_{QVL}$ is well defined over it.

The ML design values are:

- $Q \approx 311.35\,\mathrm{m}^3/\mathrm{s}$
- $V \approx 19.64 \cdot 10^6\,\mathrm{m}^3$
- $L \approx 781.30\,\mathrm{m}$

CONCLUSIONS

The univariate approach is still widely used in Earth Sciences for design purposes, for sizing works, and for risk analysis and assessment. The reasons are that

- it is simple (essentially, just invert a univariate CDF);
- the results are of immediate interpretation.

However, natural phenomena are often ruled by a combination of (non-independent) relevant variables. Considering just one variable at a time may lead to

- expensive and / or inappropriate decisions;
- under- / over-estimation of the risk.

The introduction of Copulas in Earth Sciences has made things simpler and more tractable. The possibility of exploiting the advantages of a feasible multivariate approach is no more a dream!