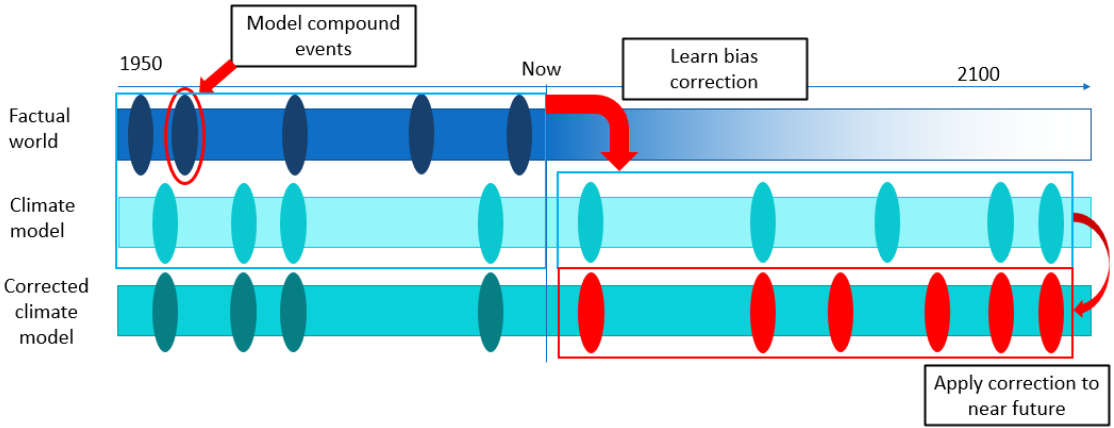
Compound climate events: can climate simulations be improved by bias correction?

Grégoire Jacquemin (Mines Paris-PSL, LSCE, INRAE), Mathieu Vrac (LSCE), Denis Allard (INRAE) & Xavier Freulon (Mines Paris-PSL)

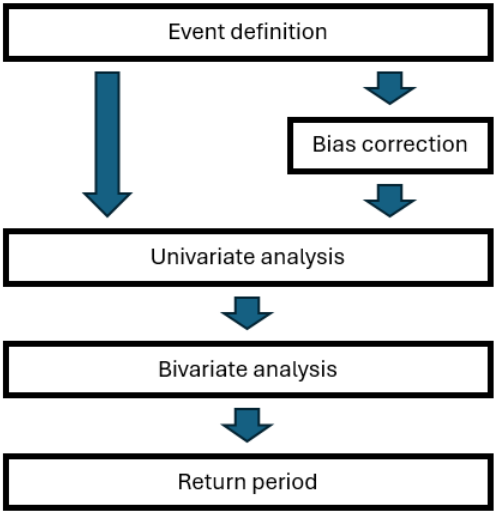
05/07/2024

Outline

Climate simulations to project compound events evolution



Key points in January



- Events definition and API
- CDF-t (univariate bias correction)
- GPD modelling and declustering
- Copula (declustering, BIC, Gumbel ...)
- Return period formulas for univariate and bivariate

Remaining questions in January

- What are the best API parameters for the two events ? A study about the correlation between the soil moisture and the API has been conducted for the German/Belgium event, but how to proceed with the second event ?
- Introduction of MBC and comparison with uncorrected data and CDF-t
- How to extend the framework to more complex events ($d > 2$)? A new methodology, Pareto processes, for bivariate modelling is proposed and will be compared to the copula approach.
- The Pareto process approach implies the use of a new univariate modelling approach, the extended GPD (EGPD) [1].
- How to account for the non-concommitance of the two univariate extremes in the bivariate return period ?

Work done since January

- Sensibility analysis on API parameters (2 weeks)
- Implementation of Multivariate Bias Correction (MBC) algorithms (3 weeks)
- Work on return periods for non-concordant events (1.5 month)
- Pareto process (1.5 month)
- Bivariate extremal index (1 week)
- English lessons and presentations (CSI and IMSC) (2 weeks)

The Seine/Loire compound event

Spatial compound event

Huge floods of Seine and Loire in June 2016 [2]

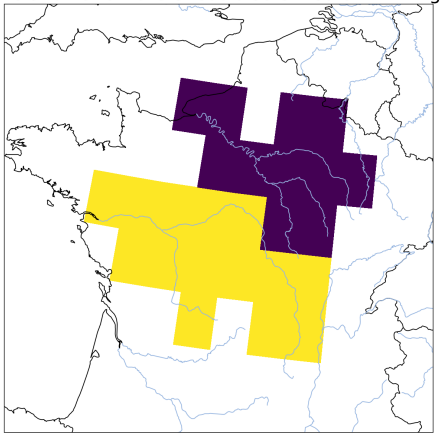
The **Antecedent Precipitation Index (API)** [3] is used to model the event:

$$API_j = \sum_{i=1}^{N} Precip_{j-i} * k^{i-1}$$

with k = 0.88 and N = 17

Daily precipitation are averaged over the Seine and the Loire watersheds for May and June between 1992 and 2021 (on ERA5 1°x1° grid)

Seine and Loire watersheds discretized on 1°x1° grid



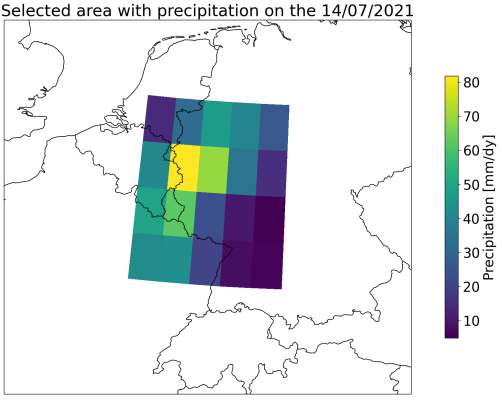
The German/Belgium compound event

Preconditioned compound event

Extremely heavy precipitation after moderate precipitation lead to a massive flood of the Ahr river in July 2021 [4]

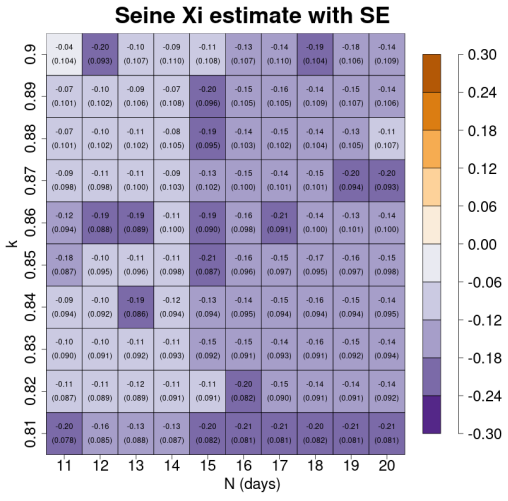
The daily precipitation (TP) and the API are used to model the event. Here the API (with k = 0.9 and N = 30) is used as a proxy for **soil moisture**

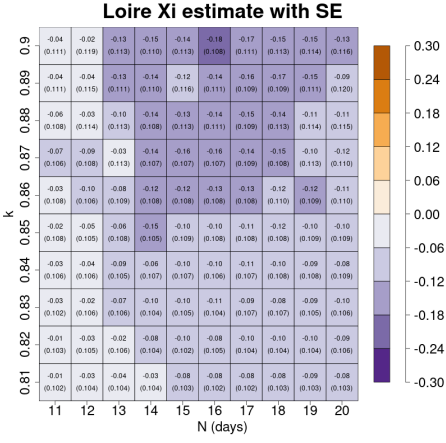
Daily precipitation are averaged over the shown area for June, July and August between 1992 and 2021 (on ERA5 1°x1° grid)



Outline

Sensibility analysis over Xi





Selection of k = 0.88 and N = 17

Outline

Data and materials

1. All the considered runs follow the **ssp5-8.5 scenario**
2. We define **4 climatic periods of 30 years** each: 1992-2021, 2022-2051, 2041-2070, 2071-2100
3. We apply bias correction algorithms on a selection of **10 GCMs**: BCC, CanESM5, CNRM-CM6, CNRM-CM6-HR, CNRM-ESM2, INM-CM4, INM-CM5, IPSL, MIROC6, MRI-ESM2
4. **6 bias correction methods** are compared: no correction, CDF-t, dOTC, R2D2 v2 (with a bivariate pivot), R2D2 with a pivot on the first variable and R2D2 with a pivot on the second variable

Multivariate bias correction algorithms

1. **R2D2 (rank resampling)**: First perform a univariate bias correction (CDF-t), then associate, in the rank space, points from the simulated data to the reference data (rank analogues) and replace the simulated values by the ones corresponding to the rank of the analogues (Vrac and Thao, 2020 [5]). This rank analogy needs a reference, which can be one variable, or several, in which case the euclidian distance is used to find the closest analogue.

2. **dOTC (optimal transport)**: multivariate optimal transport is calculated between the reference data and the model data of the historic period, and between the model data of the historic period and the projection period. This two projection plans are then applied to correct the projected data of the model (Robin et al., 2019 [6])

Outline

Models in high dimension

The copula modelling presents some interests like the decoupling of the univariate and the dependence structure, the quantification of the nature of the dependence with the copula family and the value of its parameters ... However, its generalisation in higher dimension is complex, with either multivariate copulas or vine copulas.

We propose a new modelling which shall scale to higher dimensions more easily. It keeps the decoupling of the univariate and the dependence structure and is non-parametric for the multivariate part. The drawback is that a modelling of the whole univariate distribution is needed, not only the tail.

Extended GPD

Let's consider X and Y, and note F_X and F_Y their respective cdf, modelled with the extended GPD from Naveau et al, (2016) [1].

EGPD (Naveau et al, (2016))

$$F(x) = \left(1 - \left(1 + \frac{\xi x}{\sigma}\right)^{\frac{-1}{\xi}}\right)^{\kappa}$$

Problem: in their paper, they developed the theory and the MLE estimator for $\xi > 0$. In our case, we generally have negative ξ. In Legrand et al., (2023) [7], they affirm that the theory still stands for negative ξ, but not the MLE estimator. For the moment, parameter estimation is done with the gamlss R package.

Pareto process (Delta modelling)

A transformation is applied to X and Y to get X^e and Y^e following an exponential distribution.

Let's consider a high quantile $0 < p < 1$ (for example $p = 0.95$) and we define Z_1 and Z_2 by :

$$Z_1 = X^e - F_{X^e}^{-1}(p), \quad Z_2 = Y^e - F_{Y^e}^{-1}(p))$$

According to Rootzén et al. (2018) [8], their exists a random vector $\mathbf{T} = (T_1, T_2)$ such that $\mathbf{Z} = (Z_1, Z_2)$ and $E + \mathbf{T} - \max(\mathbf{T})$ are equal in distribution, with E a unit exponential random variable, independent from $\mathbf{T}$.

Delta in Legrand et al. (2023)

We define $\Delta = Z_1 - Z_2 = T_1 - T_2$ as in Legrand et al. (2023) [7].

In their paper, the modelling was applied to wave height at different time and location, and their objective was to simulate data. With the multivariate decomposition of the extremes in E and Δ, they simply generate independent and identically distributed variables, with bootstrapping for Δ.

We propose to go further and use this modelling to compute exceedance probability.

Delta in our modelling

We suppose that Δ is continuous with density f_Δ. The decomposition of $\mathbf{Z}$ can be rewritten in terms of Δ :

$$Z_1 = E + \Delta \mathbb{1}_{\Delta<0}, \quad Z_2 = E - \Delta \mathbb{1}_{\Delta\geq 0}$$

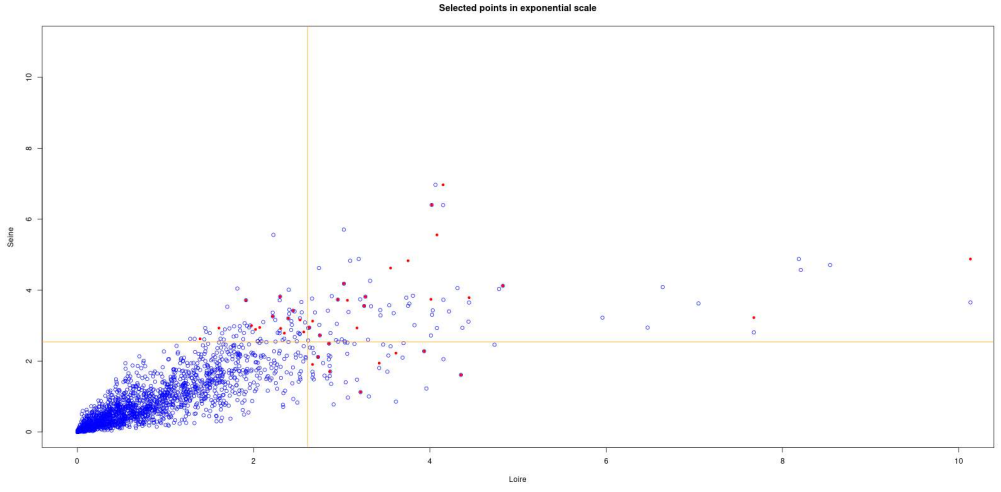
Let's consider a x and a y (they will be the return levels). We define u and v by:

$$u = -\log\left(\frac{1-F_X(x)}{p}\right), \quad v = -\log\left(\frac{1-F_Y(y)}{p}\right)$$

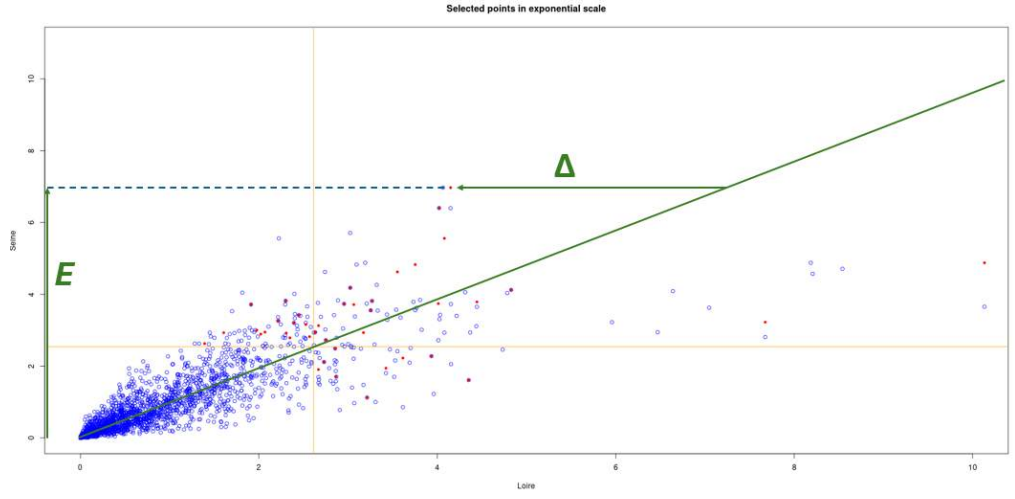
We show that :

$$\mathbb{P}\left(X > x, Y > y\right) = e^{-v}\int_{(u-v)_+}^{+\infty} e^{-t}F_\Delta(t)\,dt - e^{-u}\int_{-\infty}^{(u-v)_-} e^{t}F_\Delta(t)\,dt$$

Point selection for Delta



Point selection for Delta



Outline

Non-concomitant compound events

A compound event is defined as an exceedance over a threshold of both variables at a relatively close time. For example, the Seine/Loire event can be considered a composed event even if the floods are separated by a few days.

We define clusters and an event is the point-wise maximum of the cluster if both values also exceed their respective threshold. With this construction, we get extreme points that can be considered independent for the MLE estimation, and the probability that the point-wise maximum of a cluster is above a bivariate threshold is the probability we are looking for to get the return period.

Point process

In Resnick (1987), we found a general setting to express return periods, scale the definition with the dimensions, and the possibility to extrapolate.

We note $M_n = \max_{i \in [\![1,n]\!]}(X_i)$

Following Resnick (1987)[9], we define the two following point process:

$$\eta(x) = \inf_n (M_n > x)$$

and we have : $\{\eta(x) \leq t\} = \{M_t > x\}$

$$\mathbb{P}[\eta(x) \leq t] = \mathbb{P}[M_t > x] = 1 - \mathbb{P}[M_t \leq x] = 1 - F^{\theta t}(x)$$

Point process and return period

We define $\eta^*(x) = \inf_n \left(\max_{j \in [\![1,n]\!]}(M(J_j)) > x\right)$ with J_j being the clusters defined earlier.

We have the same results as before with r, the size of the clusters.

In the univariate setting, defining T as the return period corresponding to the return level x_T, we have : $\mathbb{P}\left[\eta^*(x_T) \leq \frac{nT}{r}\right] \simeq 1 - e^{-1}$

For the bivariate case, we can do the same : $\eta^*(x, y) = \inf_n \left(\max_{j \in [\![1,n]\!]}(\mathbf{M}(J_j)) > (x, y)\right)$

We define the bivariate return period by : $\mathbb{P}\left[\eta^*(x_T, y_T) \leq \frac{nT}{r}\right] \simeq 1 - e^{-1}$

Multivariate extremal index

In the calculus of the bivariate return period appears the bivariate extremal index, which quantifies how the data clusters in the extreme.

However, the multivariate extremal index θ is a function, which requires a more complex estimation scheme:

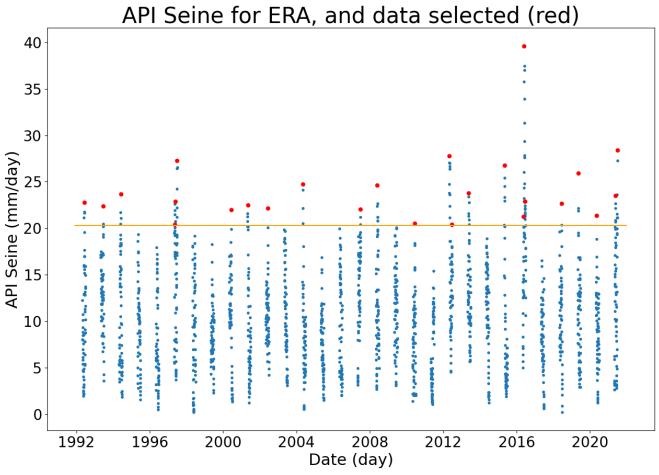
1. Reduce both variables to Fréchet margins (with the empirical cdf)
2. Select at each time step the maximum among the two values
3. Run the univariate estimator on the selected data

Univariate data and declustering

The API is strongly auto-correlated
$\Longrightarrow$ declustering

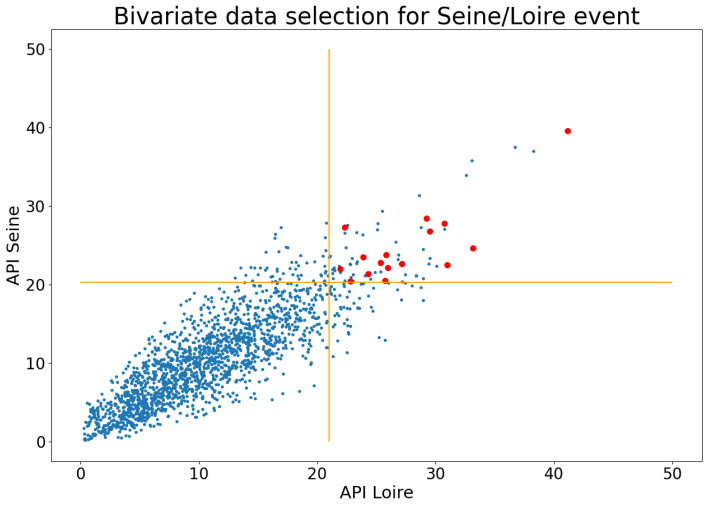
Clusters are separated by 13 days, corresponding to temporal correlation < 0.10

The **maximum of the cluster** is selected for the GPD parameter estimation, but no declustering is done for the EGPD parameter estimation



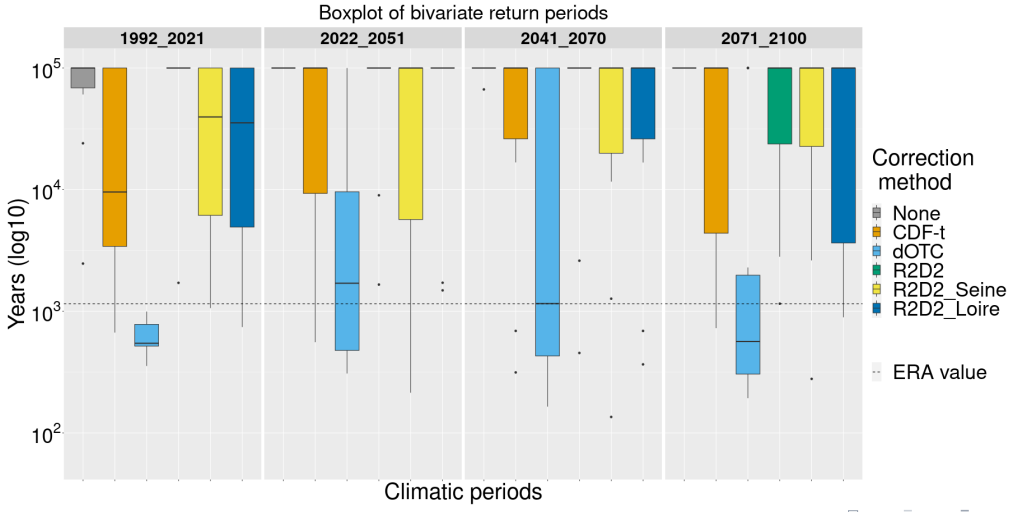
Bivariate data and copula selection

The bivariate data selection is used for both the Copula modelling and the Delta modelling.

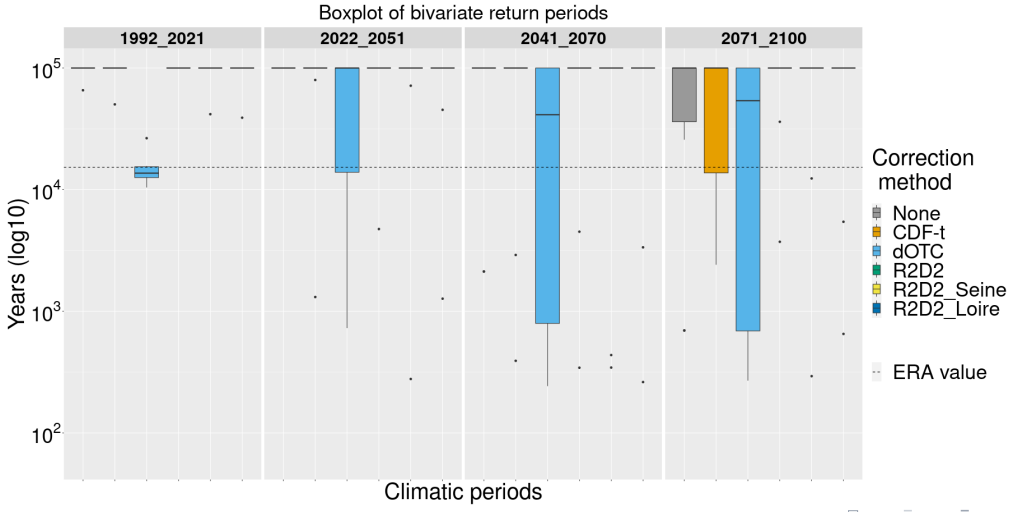


Outline

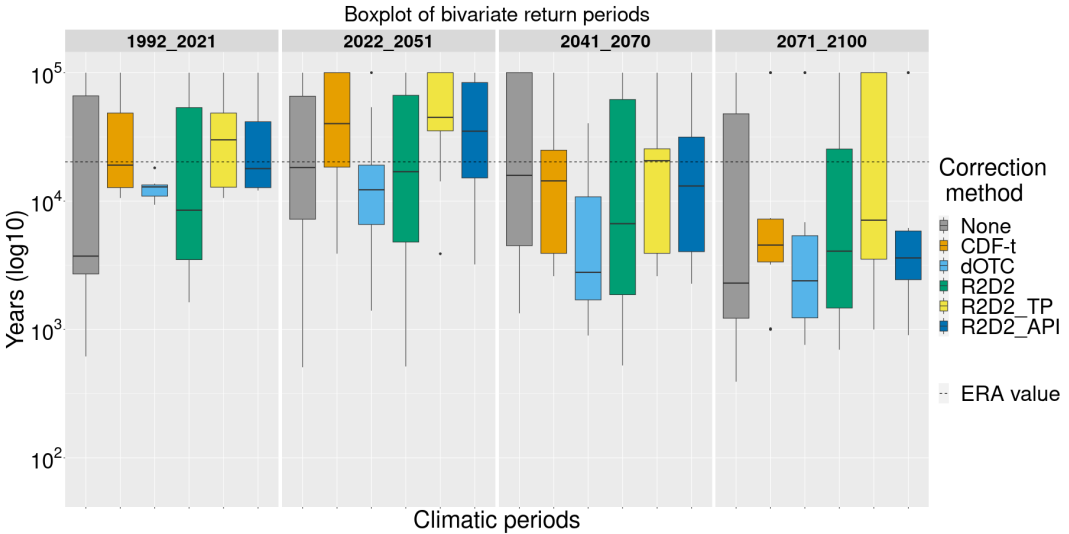
Bivariate return periods for Seine/Loire event with Copula



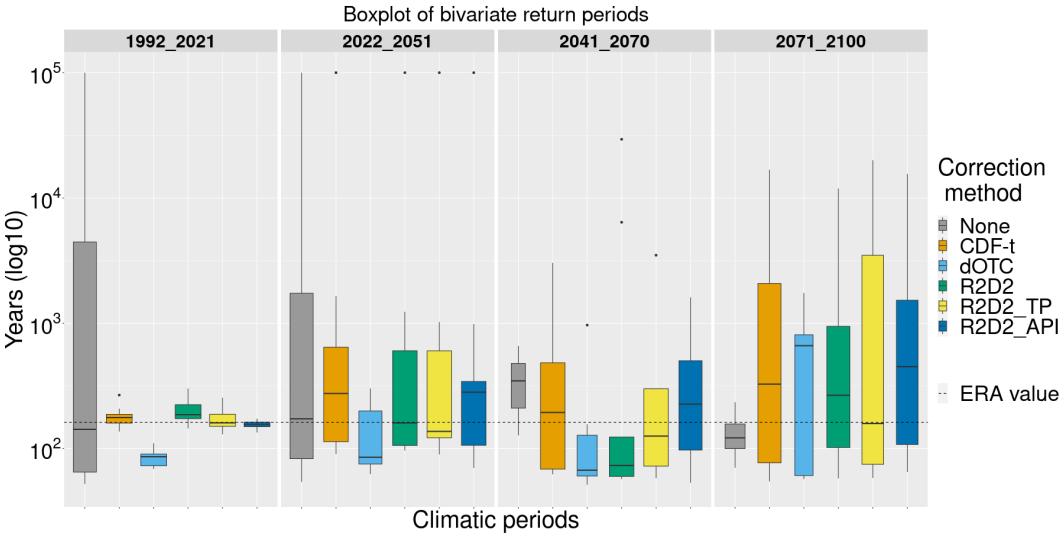
Bivariate return periods for Seine/Loire event with Delta



Bivariate return periods for German/Belgium event with Copula



Bivariate return periods for German/Belgium event with Delta



Work done since January

1. Study of API sensibility and parameters selection
2. Implementation of MBCs
3. Development of a new bivariate modelling of extremes
4. Construction of a bivariate return period formula that take into account non-concomitant compound events
5. Presentation at IMSC (International Meeting on Statistical Climatology) : https://chaire-geolearning.org/

Road map

1. Investigation of the difference in return periods for the GPD/Copula modelling and the EGPD/Delta modelling
2. Redaction of an article : "Projecting frequencies of extreme rainfall compound events under climate change using bivariate extreme value modeling and bivariate bias corrections"
3. Work on more complex compound event, like convective storms (change of resolution needed)
4. Redaction of thesis manuscript, second article (adaptation of the present framework to more complex compound events), PhD defence and international conference (EGU, EVA ...)

Bibliography I

[1] Philippe Naveau et al. "Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection". In: *Water Resources Research* 52.4 (2016), pp. 2753–2769.

[2] Susanna Mohr et al. "A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe. Part 1: Event description and analysis". In: *Natural Hazards and Earth System Sciences Discussions* 2022 (2022), pp. 1–44.

[3] Max Adam Kohler and Ray K Linsley. *Predicting the runoff from storm rainfall*. Vol. 30. US Department of Commerce, Weather Bureau, 1951.

[4] Geert Jan van Oldenborgh et al. "Rapid attribution of the May/June 2016 flood-inducing precipitation in France and Germany to climate change". In: *Hydrology and Earth System Sciences Discussions* 2016 (2016), pp. 1–23.

[5] Mathieu Vrac and Soulivanh Thao. "R 2 D 2 v2. 0: accounting for temporal dependences in multivariate bias correction via analogue rank resampling". In: *Geoscientific Model Development* 13.11 (2020), pp. 5367–5387.

Bibliography II

[6] Yoann Robin et al. "Multivariate stochastic bias corrections with optimal transport". In: *Hydrology and Earth System Sciences* 23.2 (2019), pp. 773–786.

[7] Juliette Legrand et al. "Joint stochastic simulation of extreme coastal and offshore significant wave heights". In: *The Annals of Applied Statistics* 17.4 (2023), pp. 3363–3383.

[8] Holger Rootzén, Johan Segers, and Jennifer L Wadsworth. "Multivariate generalized Pareto distributions: Parametrizations, representations, and properties". In: *Journal of Multivariate Analysis* 165 (2018), pp. 117–131.

[9] Sidney I Resnick. *Extreme values, regular variation, and point processes*. Vol. 4. Springer Science & Business Media, 1987.