Return period of non-concurrent climate compound events: a non parametric bivariate Generalized Pareto approach

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

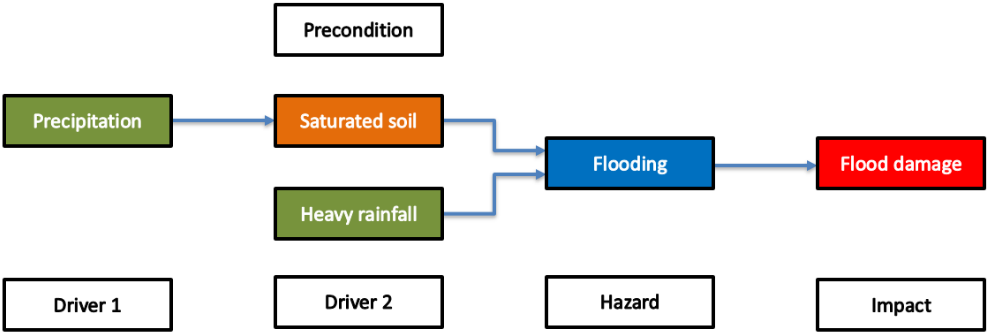
02/04/2025

Outline

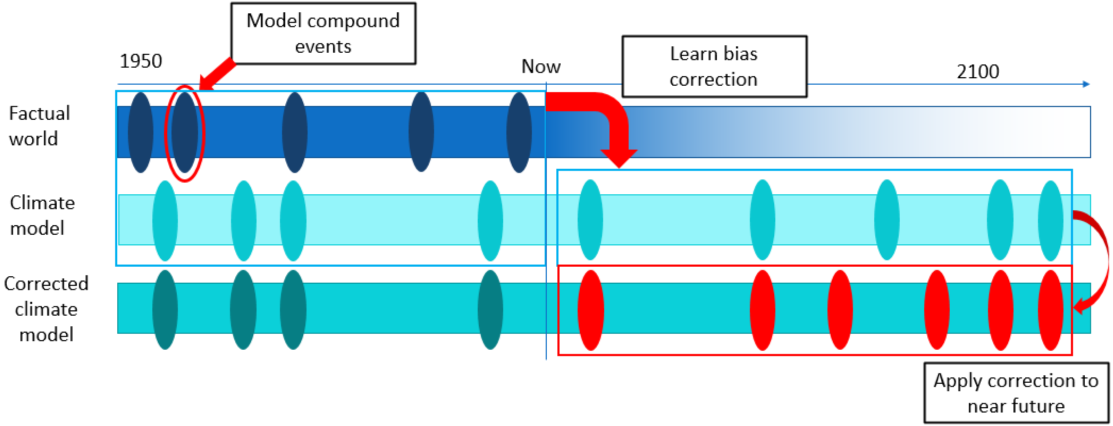
Compound event definition

Definition

A **compound event** is the combination of multiple drivers and/or hazards that contributes to societal or environmental risk (Zscheischler et al., 2020)



Projecting the evolution of the frequency of compound events



The Seine/Loire compound event

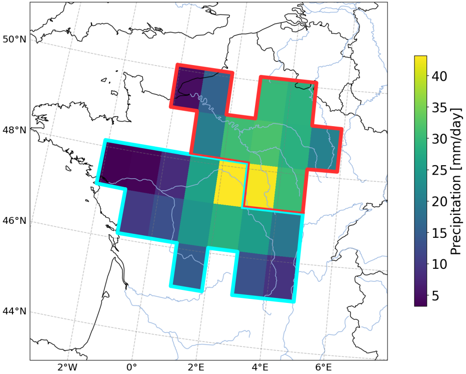
Spatial compound event Huge floods of Seine and Loire in June 2016 (Mohr et al., 2022)

The **Antecedent Precipitation Index (API)** (Kohler and Linsley, 1951) 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)

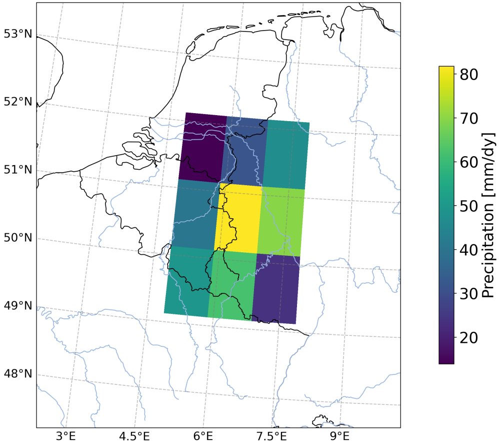


The German/Belgium compound event

Preconditioned compound event Extremely heavy precipitation after moderate precipitation lead to a massive flood of the Ahr river in July 2021 van Oldenborgh et al. (2016)

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

Generalized Pareto Distribution (GPD)

The cumulative distribution function (cdf) of a **Generalized Pareto Distribution (GPD)** with location $\mu \in \mathbb{R}$, scale $\sigma > 0$, and shape $\xi \in \mathbb{R}$ is defined as:

Cumulative distribution function of the GPD

$$G_{\xi,\mu,\sigma}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\frac{x-\mu}{\sigma}\right) & \text{if } \xi = 0, \end{cases}$$

where $x \geq \mu$

Extended Generalized Pareto Distribution (EGPD)

The **Extended Generalized Pareto Distribution (EGPD)** (Naveau et al., 2016) allows a complete modeling of the distribution, eg:

Cumulative distribution function of the EGPD

$f(G(x|\xi,\sigma)) = G(x|\xi,\sigma)^{\kappa}$, with $\kappa > 0$

where G is the GPD cumulative distribution function ($\xi > -0.5$).

- There exists other forms for f.

Extremal index

As the API is by construction a strongly auto-correlated variable, declustering is needed and the strength of this auto-correlation can be represented by the **extremal index θ**:

- Consider an i.i.d. sample $\tilde{X}_i$ following the same distribution as X_i.
- $M_n = \max_i(X_i)$ with CDF F_{max} and $\tilde{M}_n = \max_i(\tilde{X}_i)$ with CDF $\tilde{F}_{max}$.
- The extremal index is the real number $0 < \theta \leq 1$

$$F_{max} = \tilde{F}_{max}^{\theta}.$$

- It is estimated in practice with the Dgaps algorithm (Holešovský and Fusek, 2020)
- The extremal index can be extended to higher dimensions

Extremal index

As the API is by construction a strongly auto-correlated variable, declustering is needed and the strength of this auto-correlation can be represented by the **extremal index θ**:

- Consider an i.i.d. sample $\tilde{X}_i$ following the same distribution as X_i.
- $M_n = \max_i(X_i)$ with CDF F_{max} and $\tilde{M}_n = \max_i(\tilde{X}_i)$ with CDF $\tilde{F}_{max}$.
- The extremal index is the real number $0 < \theta \leq 1$

$$F_{max} = \tilde{F}_{max}^{\theta}.$$

- It is estimated in practice with the Dgaps algorithm (Holešovskỳ and Fusek, 2020)
- The extremal index can be extended to higher dimensions

Extremal index

As the API is by construction a strongly auto-correlated variable, declustering is needed and the strength of this auto-correlation can be represented by the **extremal index θ**:

- Consider an i.i.d. sample $\tilde{X}_i$ following the same distribution as X_i.
- $M_n = \max_i(X_i)$ with CDF F_{max} and $\tilde{M}_n = \max_i(\tilde{X}_i)$ with CDF $\tilde{F}_{max}$.
- The extremal index is the real number $0 < \theta \leq 1$

$$F_{max} = \tilde{F}_{max}^{\theta}.$$

- It is estimated in practice with the Dgaps algorithm (Holešovskỳ and Fusek, 2020)
- The extremal index can be extended to higher dimensions

Extremal index

As the API is by construction a strongly auto-correlated variable, declustering is needed and the strength of this auto-correlation can be represented by the **extremal index θ**:

- Consider an i.i.d. sample $\tilde{X}_i$ following the same distribution as X_i.
- $M_n = \max_i(X_i)$ with CDF F_{max} and $\tilde{M}_n = \max_i(\tilde{X}_i)$ with CDF $\tilde{F}_{max}$.
- The extremal index is the real number $0 < \theta \leq 1$

$$F_{max} = \tilde{F}_{max}^{\theta}.$$

- It is estimated in practice with the Dgaps algorithm (Holešovskỳ and Fusek, 2020)
- The extremal index can be extended to higher dimensions

Extremal index

As the API is by construction a strongly auto-correlated variable, declustering is needed and the strength of this auto-correlation can be represented by the **extremal index θ**:

- Consider an i.i.d. sample $\tilde{X}_i$ following the same distribution as X_i.
- $M_n = \max_i(X_i)$ with CDF F_{max} and $\tilde{M}_n = \max_i(\tilde{X}_i)$ with CDF $\tilde{F}_{max}$.
- The extremal index is the real number $0 < \theta \leq 1$

$$F_{max} = \tilde{F}_{max}^{\theta}.$$

- It is estimated in practice with the Dgaps algorithm (Holešovskỳ and Fusek, 2020)
- The extremal index can be extended to higher dimensions

Extremal index

As the API is by construction a strongly auto-correlated variable, declustering is needed and the strength of this auto-correlation can be represented by the **extremal index θ**:

- Consider an i.i.d. sample $\tilde{X}_i$ following the same distribution as X_i.
- $M_n = \max_i(X_i)$ with CDF F_{max} and $\tilde{M}_n = \max_i(\tilde{X}_i)$ with CDF $\tilde{F}_{max}$.
- The extremal index is the real number $0 < \theta \leq 1$

$$F_{max} = \tilde{F}_{max}^{\theta}.$$

- It is estimated in practice with the Dgaps algorithm (Holešovskỳ and Fusek, 2020)
- The extremal index can be extended to higher dimensions

Return period of auto-correlated variable

Return level and return period

The return period T is the expected waiting time between two exceedances above a "return level" x_T.

Considering that $F_{\text{max}}(x) = \mathbb{P}(\max_i(X_i) \leq x) \simeq F^{N\theta}(x)$, with N being the number of X_i per year, one thus gets:

$$T = \frac{1}{1 - F_{\text{max}}(x_T)} \simeq \frac{1}{1 - F^{N\theta}(x_T)} \Leftrightarrow T \simeq \frac{1}{N\theta\mathbb{P}\left(X > x_T\right)},$$

where θ is the extremal index.

Return period of auto-correlated variable

Return level and return period

The return period T is the expected waiting time between two exceedances above a "return level" x_T.

Considering that $F_{\max}(x) = \mathbb{P}(\max_i(X_i) \leq x) \simeq F^{N\theta}(x)$, with N being the number of X_i per year, one thus gets:

$$T = \frac{1}{1 - F_{\max}(x_T)} \simeq \frac{1}{1 - F^{N\theta}(x_T)} \Leftrightarrow T \simeq \frac{1}{N\theta\mathbb{P}\left(X > x_T\right)}, \tag{1}$$

where θ is the extremal index.

Outline

Copula modeling for extreme values

Theorem (Sklar, 1959)

Let F be the multivariate cumulative distribution function of a random vector (X_1, X_2). Then there exists a function $C : [0,1]^2 \to [0,1]$ called a **copula** defined by:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)).$$

If the $F_1, ..., F_d$ are continuous, the copula C is unique.

This allows us to propose the following approach for extreme values:

1. Propose a univariate extreme model for the marginals (GPD)
2. Reduce the tail of the distributions to uniform margins
3. Determine the copula

Copula modeling for extreme values

Theorem (Sklar, 1959)

Let F be the multivariate cumulative distribution function of a random vector (X_1, X_2). Then there exists a function $C : [0,1]^2 \to [0,1]$ called a **copula** defined by:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)).$$

If the $F_1, ..., F_d$ are continuous, the copula C is unique.

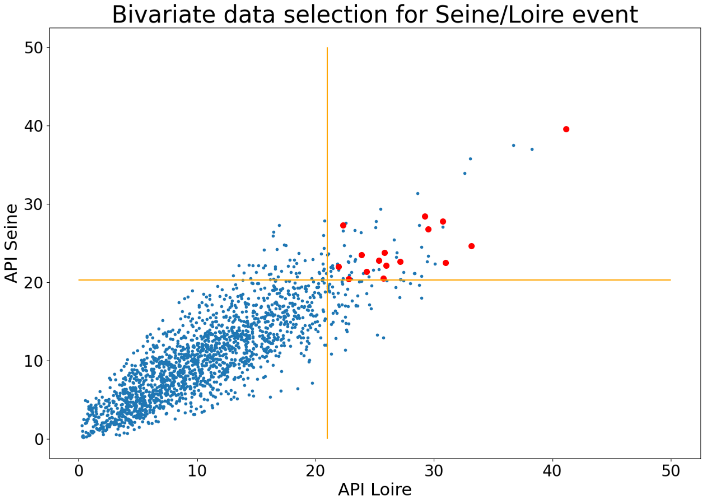
This allows us to propose the following approach **for extreme values**:

1. Propose a univariate extreme model for the marginals (GPD)
2. Reduce the tail of the distributions to uniform margins
3. Determine the copula

Bivariate data and copula selection

The GPD and copula parameters are estimated with maximum likelihood.

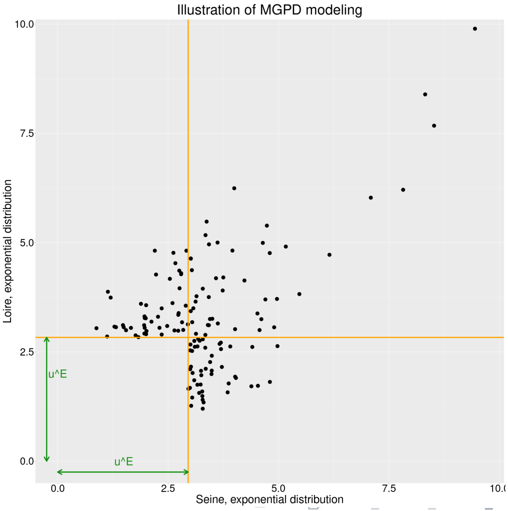
The best copula is selected among a few parametric families according to the Bayesian Information Criteria (BIC).



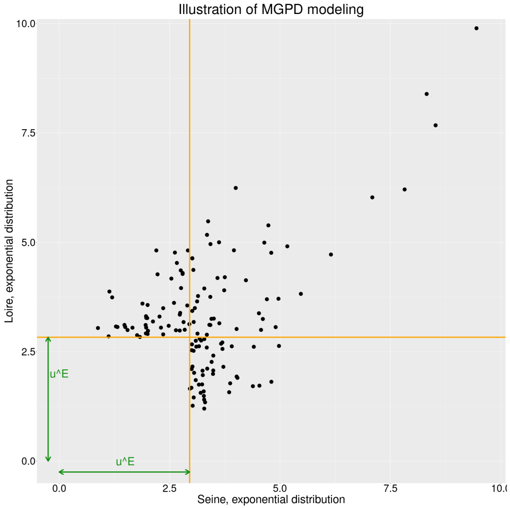
Multivariate Generalized Pareto Distribution (MGPD)

- The Extended Generalized Pareto Distribution (EGPD) is used to model the complete distribution.
- Thanks to the EGPD, one can construct $\boldsymbol{X}^E$ the exponential transform of $\boldsymbol{X}$ and define for $\boldsymbol{u}^E$ a high enough threshold:

$$\boldsymbol{Z} := \boldsymbol{X}^E - \boldsymbol{u}^E \mid \boldsymbol{X}^E \nleq \boldsymbol{u}^E \sim MGPD(\mathbf{0}, \mathbf{1})$$



Multivariate Generalized Pareto Distribution (MGPD)



- Rootzén et al. (2018) established that, if $\boldsymbol{Z}$ follows an unitary MGPD, their exists a random vector $\boldsymbol{T}$ such that:

$$\boldsymbol{Z} \overset{\text{law}}{=} E + \boldsymbol{T} - \max(\boldsymbol{T})$$

where E follows an unitary exponential distribution, independent from $\boldsymbol{T}$.

Delta decomposition for bivariate GPD (biGPD)

In a bivariate context, Legrand et al. (2023) defined the random variable $\Delta = Z_1 - Z_2 = T_1 - T_2$.

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

For $x_1, x_2 \geq u_1, u_2$, $\mathbb{P}\left(X_1 > x_1, X_2 > x_2\right)$ can be expressed using the empirical CDF of Δ and numerical integration.

Delta decomposition for bivariate GPD (biGPD)

In a bivariate context, Legrand et al. (2023) defined the random variable $\Delta = Z_1 - Z_2 = T_1 - T_2$.

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

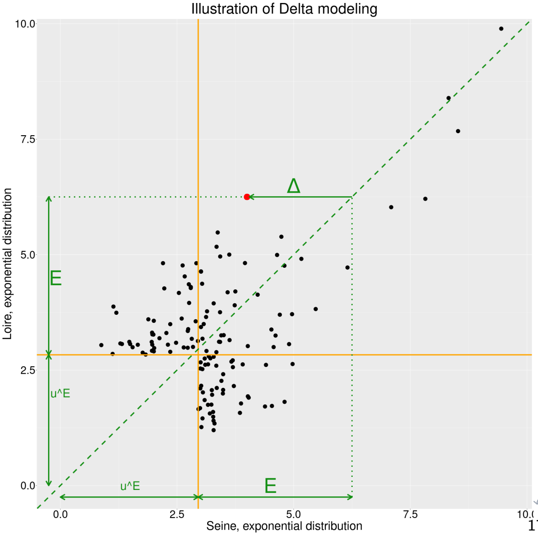
For $x_1, x_2 \geq u_1, u_2$, $\mathbb{P}\left(X_1 > x_1, X_2 > x_2\right)$ can be expressed using the empirical CDF of Δ and numerical integration.

Delta decomposition for bivariate GPD (biGPD)

In a bivariate context, Legrand et al. (2023) defined the random variable $\Delta = Z_1 - Z_2 = T_1 - T_2$.

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

For $x_1, x_2 \geq u_1, u_2$, $\mathbb{P}(X_1 > x_1, X_2 > x_2)$ can be expressed using the empirical CDF of Δ and numerical integration.

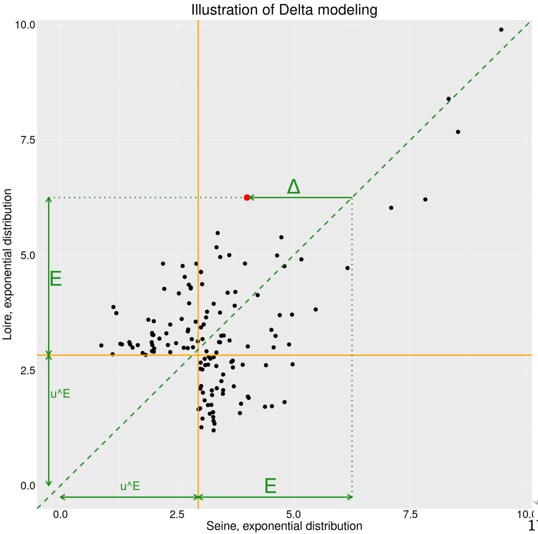


Delta decomposition for bivariate GPD (biGPD)

In a bivariate context, Legrand et al. (2023) defined the random variable $\Delta = Z_1 - Z_2 = T_1 - T_2$.

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

For $x_1, x_2 \geq u_1, u_2$, $\mathbb{P}(X_1 > x_1, X_2 > x_2)$ can be expressed using the empirical CDF of Δ and numerical integration.

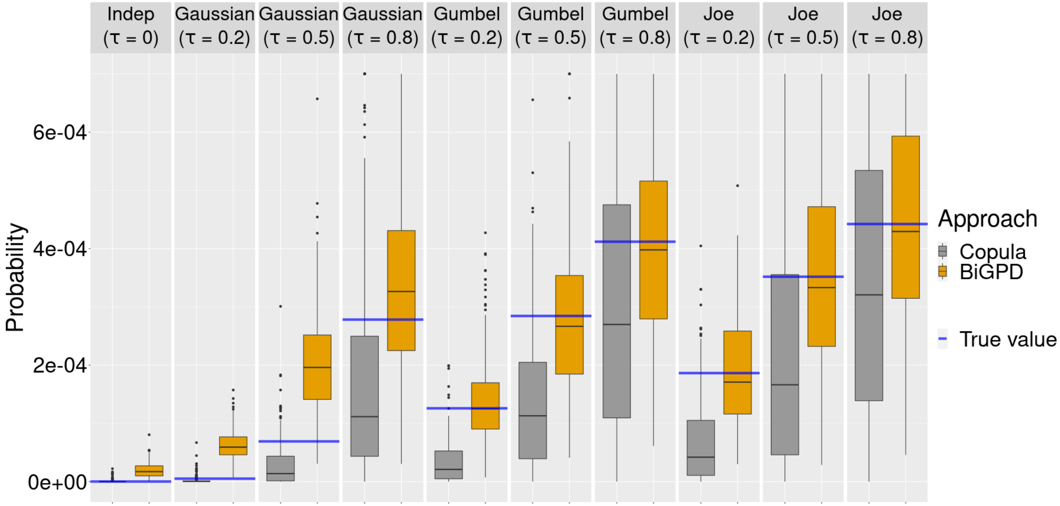


Protocol for statistical simulations

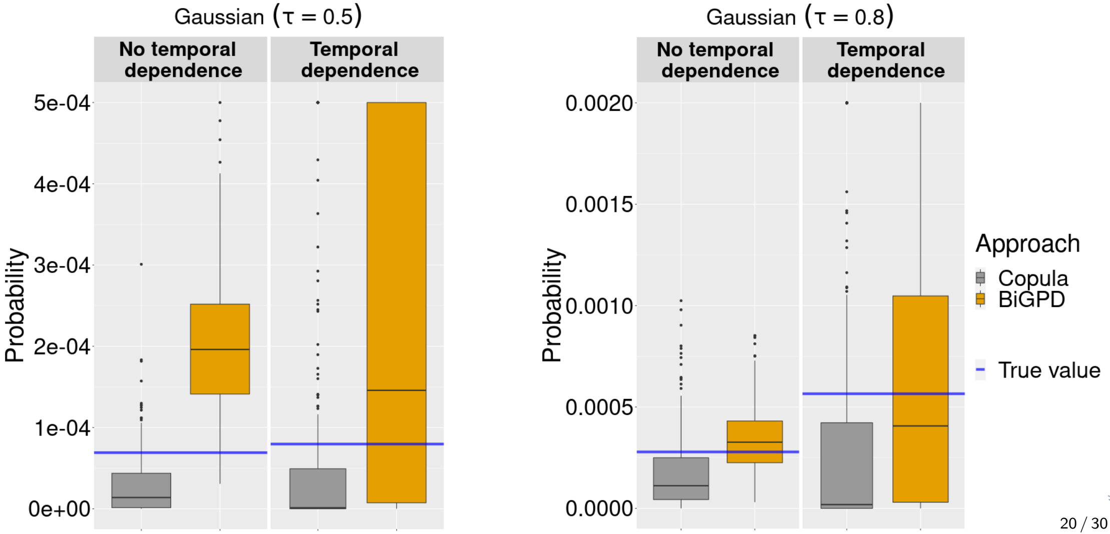
Objective: compare the two approaches (copula and biGPD) $\Longrightarrow$ statistical simulations

- 30 x 61 points are simulated following a copula (Gaussian, Gumbel or Joe)
- The univariate distribution are transformed to EGPDs
- Return periods and associated probabilities are calculated with the two approaches
- 150 draws are performed $\Longrightarrow$ boxplots to represent the uncertainties
- Exact values are known and represented by dashed lines

Statistical simulations without temporal correlation



Statistical simulations with temporal correlation

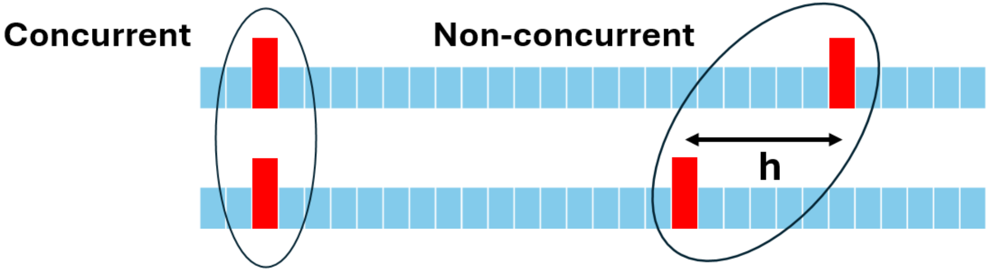


Outline

Our new contribution: non-concurrent compound event

Definition

non-concurrent compound event: a compound event where the composing variables reach extreme values not necessarily at the same time step, within a defined range.



Bivariate return period for non-concurrent compound events

$\eta(x_1, x_2)$ gives the number of the first block of size h where both x_1 and x_2 are exceeded.

Definition

The **non-concurrent bivariate return period T** associated to the return level vector $(x_{1,T}, x_{2,T})$ is such that:

$$\mathbb{P}\left[\eta(x_{1,T}, x_{2,T}) \leq \frac{NT}{h}\right] := 1 - e^{-1} \simeq 0.63. \tag{2}$$

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

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

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

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)** (Vrac and Thao, 2020):
   - Univariate bias correction (CDF-t)
   - Rank analogues: associate, in the rank space, points from the simulated data to the reference data
   - Replace the simulated values by the ones corresponding to the rank of the analogues
   - The reference for the rank analogy (the pivot) can be one variable or several

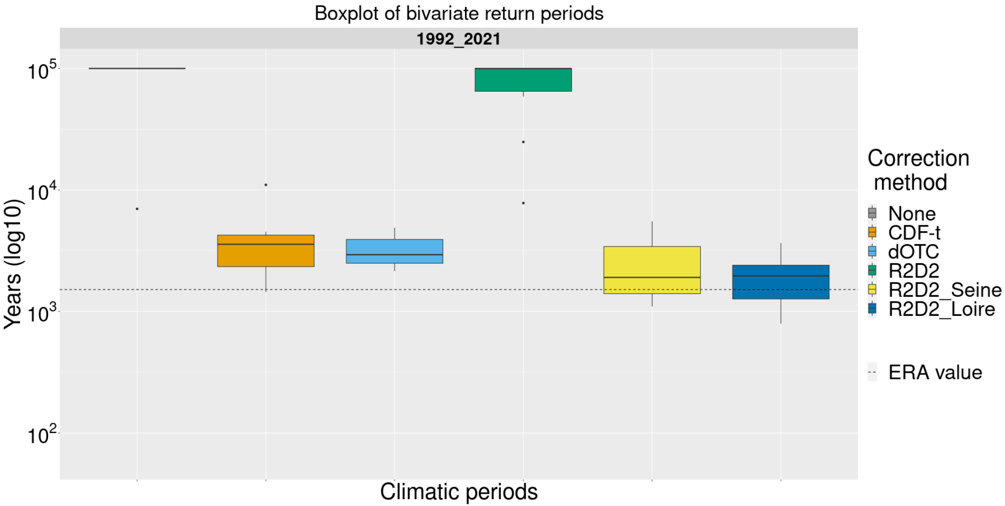
2. **dOTC (optimal transport)** (Robin et al., 2019):
   - Multivariate optimal transport between the reference data and the model data of the reference period
   - Multivariate optimal transport between the model data of the reference period and the projection period
   - The two projection plans are combined to correct the projected data of the model

Multivariate bias correction algorithms

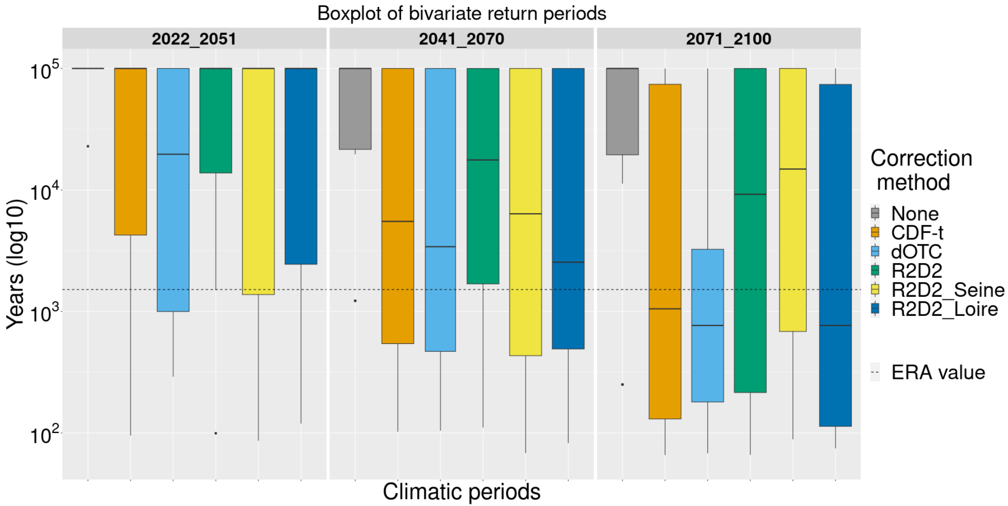
1. **R2D2 (rank resampling)** (Vrac and Thao, 2020):
   - Univariate bias correction (CDF-t)
   - Rank analogues: associate, in the rank space, points from the simulated data to the reference data
   - Replace the simulated values by the ones corresponding to the rank of the analogues
   - The reference for the rank analogy (the pivot) can be one variable or several

2. **dOTC (optimal transport)** (Robin et al., 2019):
   - Multivariate optimal transport between the reference data and the model data of the reference period
   - Multivariate optimal transport between the model data of the reference period and the projection period
   - The two projection plans are combined to correct the projected data of the model

Bivariate return periods for Seine/Loire event with biGPD



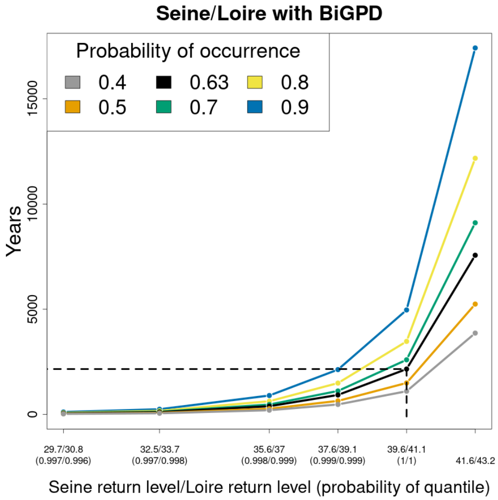
Bivariate return periods for Seine/Loire event with biGPD



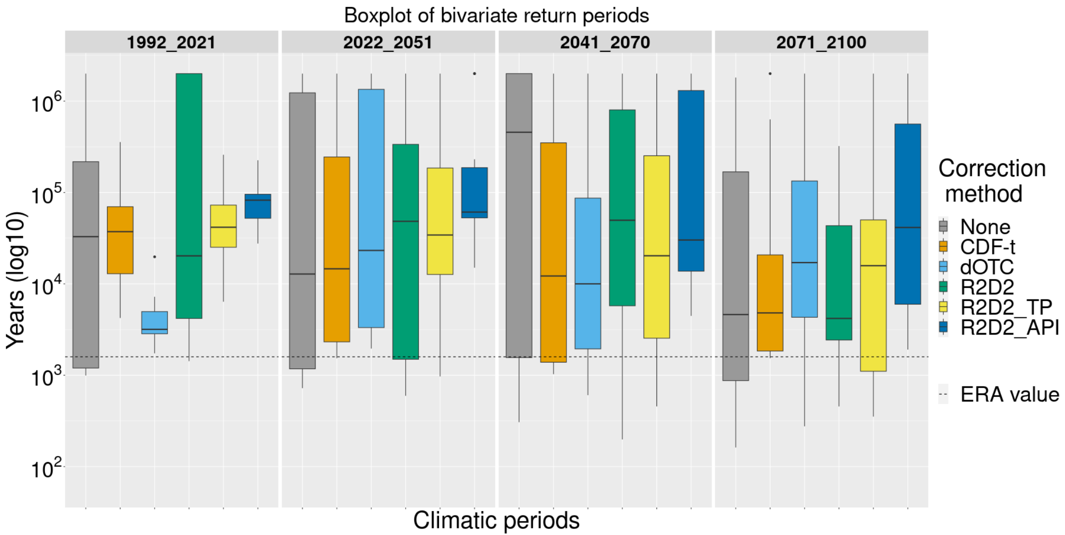
Bivariate IDF-like curve for Seine/Loire event with biGPD

IDF-like function between a value x, a time t and the probability p to exceed this value before this time:

$$p(x,t) = 1 - F^{tN\theta}(x).$$



Bivariate return periods for Germany/Belgium event with copula



Conclusion

- Fast and efficient framework to project frequency evolution of compound rain events
- New contributions:
  - Non-parametric modeling with bivariate GPD
  - Definition of non-concurrence and analytic formulas for the return periods
- Climate models have shown statistical biases on extreme events $\Longrightarrow$ bias correction necessary
- Analysis of the bias correction methods and their variability $\Longrightarrow$ starting soon
- Application of the framework to other types of compound events is planned

Chair Geolearning

Thanks to the Geolearning chair and its partners
Link to the chairs's website: https://chaire-geolearning.org/

References I

Holešovskỳ, J. and Fusek, M. (2020). Estimation of the extremal index using censored distributions. *Extremes*, 23(2):197–213.

Kohler, M. A. and Linsley, R. K. (1951). *Predicting the runoff from storm rainfall*, volume 30. US Department of Commerce, Weather Bureau.

Legrand, J., Ailliot, P., Naveau, P., and Raillard, N. (2023). Joint stochastic simulation of extreme coastal and offshore significant wave heights. *The Annals of Applied Statistics*, 17(4):3363–3383.

Mohr, S., Ehret, U., Kunz, M., Ludwig, P., Caldas-Alvarez, A., Daniell, J. E., Ehmele, F., Feldmann, H., Franca, M. J., Gattke, C., et al. (2022). A multi-disciplinary analysis of the exceptional flood event of july 2021 in central europe. part 1: Event description and analysis. *Natural Hazards and Earth System Sciences Discussions*, 2022:1–44.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Robin, Y., Vrac, M., Naveau, P., and Yiou, P. (2019). Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences*, 23(2):773–786.

References II

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Multivariate generalized pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP*, 8(3):229–231.

van Oldenborgh, G. J., Philip, S., Aalbers, E., Vautard, R., Otto, F., Haustein, K., Habets, F., Singh, R., and Cullen, H. (2016). Rapid attribution of the may/june 2016 flood-inducing precipitation in france and germany to climate change. *Hydrology and Earth System Sciences Discussions*, 2016:1–23.

Vrac, M. and Thao, S. (2020). R 2 d 2 v2. 0: accounting for temporal dependences in multivariate bias correction via analogue rank resampling. *Geoscientific Model Development*, 13(11):5367–5387.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., et al. (2020). A typology of compound weather and climate events. *Nature reviews earth & environment*, 1(7):333–347.