Journée Geolearning avec les élèves Mines Paris 3A de l'option

Géostatistique et Probabilités Appliquées

Date : **13 novembre 2023**

Lieu : **Ecole des Mines de Paris, 60 Bd Saint-Michel, Paris 6ième, salle L213**

Objectif : créer du dialogue entre les entreprises mécènes de la chaire et les élèves. Organisation en deux temps : matinée consacrée à une présentation générale des entreprises ; après-midi

09:00 Accueil et ouverture

09:10 Andra

09:50 BNP

10:30 *Pause*

10:50 SCOR

11:30 CCR

12:10 *Pause déjeuner en V117*

14:00 Exposés de doctorants : Grégoire Jacquemin, Ferdinand Bhavsar

15:15 Témoignage Antoine Doizé

15:30 *Pause*

16:00 Présentation des sujets de stage (2), de thèse (1) et de post-doc (2) par les représentants de la chaire : 10 minutes de présentation + 5 minutes discussion

17:30 Fin

GEOLEARNING
CHAIRE /// Data Science for the Environment

Séminaire Geolearning : mécènes et chercheurs

Date : **14 novembre 2023**

Lieu : **Ecole des Mines de Paris, 60 Bd Saint-Michel, Paris 6ième, salle L218**

Objectif : échanger sur les travaux en cours ; creuser les pistes scientifiques ensemble. Comme il s'agit d'un premier séminaire, on privilégie un cadre restreint. Echanger sur la construction du mastère spécialisé

09:00 Accueil et ouverture

09:10 Quels apports du Machine Learning et des Deep generative models pour la chaire Geolearning ?

10:00 Quels liens entre événements extrêmes et statistiques spatio-temporelles ?

10:50 *Pause*

11:10 Projet de mastère spécialisé

Emilie

12:10 *Pause déjeuner à l'espace Vendôme*

14:00 En lien avec la chaire Geolearning, quel seraient vos **deux verrous scientifiques majeurs** ? [On évite une longue liste ; on évite les verrous techniques/logiciels/opérationnels]

16:00 Discussion générale pour reboucler sur la journée entière. Quelles pistes prioritaires ?

GEOLEARNING
CHAIRE /// Data Science for the Environment

Témoignage d'un doctorant en première année

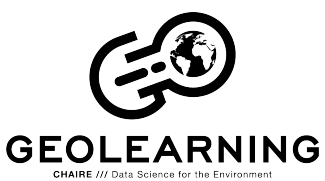
Journée Geolearning

Antoine Doizé

Advisors : D. Allard, P. Naveau, O. Wintenberger

Sorbonne-Université - Fondation des Mines de Paris

November 17, 2023



Presentation du profil

Une formation entre ingénierie et recherche 1/2

Ecole polytechnique : Département Data Science et IA (promo X17)

1. Stage Nephelaï : Algorithmes de détection d'anomalies dans des transactions financières (Classification, Modèles statistiques, Séries temporelles).
2. Stage Aqemia : Algorithmes génératifs de molécules à potentiel thérapeutique (Modèles génératifs et prédictifs, optimisation en grande dimension).

ENS Paris-Saclay : Master Mathématiques Vision Apprentissage

1. Master de recherche en IA : Deep Learning, Computational Statistics, Object Recognition, Convex Optimisation, Time Series, Bayesian Machine Learning, Kernel Methods...
2. Stage BCG Gamma : Modélisation bayésienne de l'impact des campagnes marketing. (Regression linéaire bayésienne, Chaînes de Markov, Optimisation).

Une formation entre ingénierie et recherche 2/2

Sorbonne-Université : Master Probabilité Modèles Aléatoires

1. Master de recherche en probabilités : Calcul Stochastique, Théorèmes limites, Processus de Markov, Statistiques de l'apprentissage, Géométrie Aléatoire, Bayesien Non-Paramétrique...

Stages de recherche

1. Stage de recherche Sorbonne-Université : Extreme Value Theory. Etude des propriétés du bootstrap pour les estimateurs de shape de queue de distribution.
2. Stage Inria : Extreme Value Theory pour l'étude des pluies. Modélisation du comportement des régimes extrêmes de pluie (sécheresse et pluies intenses) mesurés en une station météo.

Modèles statistiques spatiotemporels pour simuler les périodes de sécheresses et de pluies intenses sur des périodes longues et à l'échelle régionale (1/2)

Notions et difficultés clés de l'étude

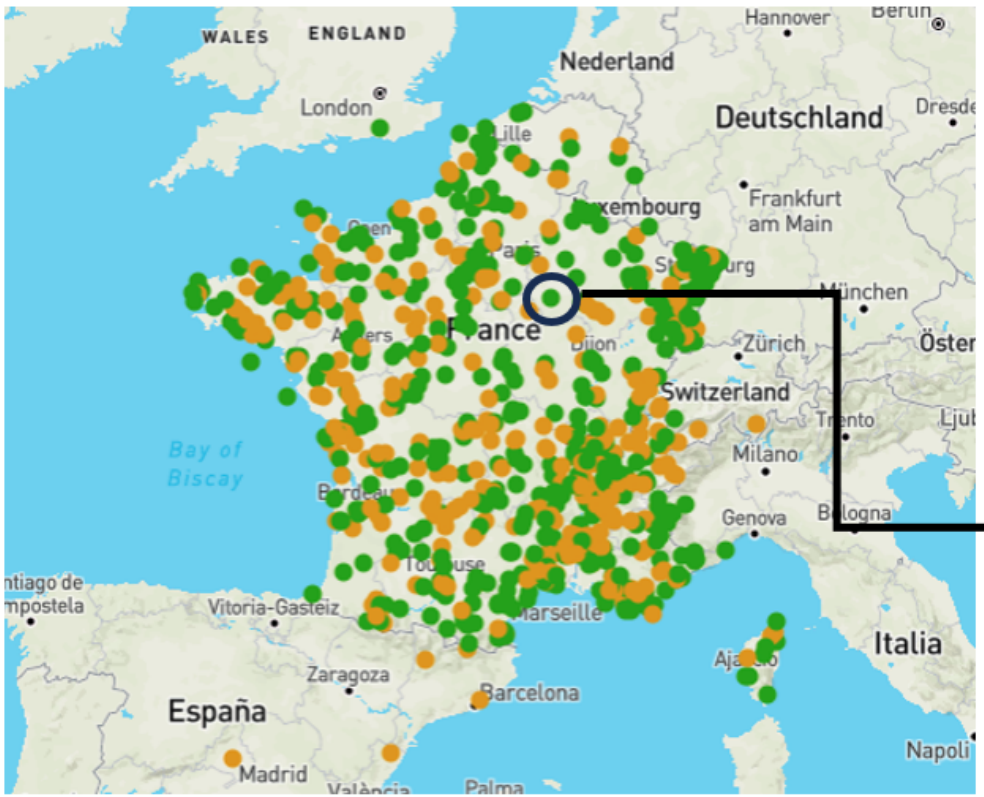
1. Etude de séries temporelles longues et sparse : il ne "pleut pas" beaucoup plus souvent qu'il ne pleut.
2. Comportements extrêmes des précipitations : On cherche particulièrement à bien modéliser les régimes extrêmes (inondations, longues sécheresses) : alors qu'on en observe peu (ils sont par définition rares).
3. Système non Markovien : il y a des phénomènes de persistance dans les sécheresses.
4. Système non stationnaire : le changement climatique impacte le comportement des précipitations (et en particulier des extrêmes).

Modèles statistiques spatiotemporels pour simuler les périodes de sécheresses et de pluies intenses sur des périodes longues et à l'échelle régionale (2/2)

Enjeux de l'étude

1. Enjeu risque climatique : mieux comprendre l'évolution des régimes des catastrophes naturelles (pluies intenses / longues sécheresses)
2. Enjeu d'aménagement du territoire :
   - Construction et phénomène de rétractation des argiles
   - Choix des cultures à implanter / des essences d'arbres les plus adaptées à un climat

Données à disposition



- précipitations journalières
- durée de plusieurs années
- mesurées sur plusieurs stations météo

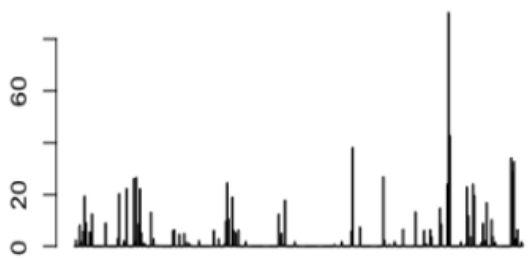


Figure: Données pluviomètres météo France

Objectif : Avoir un modèle génératif de données de pluie

Etape 1 : Proposer un modèle paramétrique qui génère des données de pluie

1. Modèles Markoviens censurés
2. Modèles semi-markoviens

$$(x_1....x_n) \sim G_\phi$$

Modifier les paramètres changera les propriétés de la série générée.

Etape 2 :Inférer les paramètres correspondant aux mieux aux données observées

On note L_ϕ la likelihood de la série générée par le générateur, et $(x_1^*...x_n^*)$ les données de pluie observées, on cherche donc

$$\phi^* = argmax_\phi L_\phi((x_1^*...x_n^*))$$

Etape 3 : On peut alors générer des données de pluie similaires aux données observées

1. Etudier les régimes de pluie
2. Etudier la dépendance de ces régimes avec des covariables climatiques (température moyenne terrestre, température à la surface des océans...)

Approche entamée : Modèles markoviens censurés

On a commencé par utiliser des modèles de Markov censurés

Exemple 1

Exemple de modèle : Censored Stochastic Reccurent Equation

$$X_t = [A^{\xi}_{1,\sigma_1,t}\mathbb{1}(X_{t-1} > 0) + A^{\xi}_{2,\sigma_2,t}\mathbb{1}(*X_{t-1} > 0)] \times X_{t-1} + B_t$$

$$\tilde{X}_t = \begin{cases} 0 \text{ if } X_t \leq 0 \\ X_t \text{ if } X_t > 0 \end{cases} .$$

- $A_{1,\sigma_1,t}$ follows a log-normal distribution
- $B_t = +/-1$ avec probabilité $\frac{1}{2}$

Difficulté : estimation de la likelihood en cadre censuré

$$X_t = [A^{\xi}_{1,\sigma,t}\mathbb{1}(X_{t-1} > 0) + A^{\xi}_{2,\sigma',t}\mathbb{1}(*X_{t-1} > 0)] \times X_{t-1} + B_t$$

$$\tilde{X}_t = \begin{cases} 0 \text{ if } X_t \leq 0 \\ X_t \text{ if } X_t > 0 \end{cases} .$$

On note $f(x_1 \dots x_n|\phi)$ avec $\phi = (\xi, \sigma, \sigma')$ la vraissemblance de la chaîne non observée $(X_1 \dots X_n)$.Son estimation est rendue difficile par la censure.

Cadre non-censuré

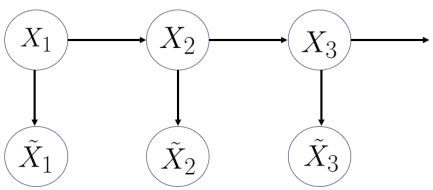


Figure: HMM Non censuré

$f(x_1, x_2, x_3|\phi)$

Cadre censuré simple

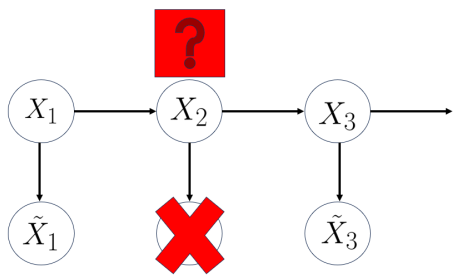


Figure: HMM censuré

$\int_{-\infty}^{0} f(x_1, x_2, x_3|\phi)dx_2$

Cas général En notant $c_t = (x_1 \dots x_t)$ la totalité de la chaîne et $c_t^{\leq 0}$ le vecteur de variables censurées

$$\int_{-\infty}^{0} f\left(\mathbf{c}_t \mid \boldsymbol{\phi}\right) \mathrm{d}\mathbf{c}_t^{\leq 0}$$

Cette intégrale intractable quand il y a beaucoup de censure, ce qui est notre cas.

Pistes de solution : estimation de la likelihood en cadre censuré

1. **Expectancy-Maximisation algorithm.** Essayer successivement d'estimer et de maximiser de façon approximative la likelihood. Méthode rapidement limitée.
2. **Pairwise Composite likelihood methods.** Utiliser une approximation de la likelihood. Celle-ci ne travaille pas sur la totalité de la chaîne mais sur des "paires" successives d'observations : $\mathcal{L}_{C,pair} = \prod_{t=1}^{n} f_{\alpha,\alpha',\sigma,\sigma'}(x_t, x_{t+1})$
3. **Méthodes likelihood-free.** Utiliser des méthodes "black-box" qui font une estimation des paramètres en fonction de la chaîne. (Deep Learning : Multilayer Percepron, Convolutionnal Neural Networks, Long-Short-Term-Memory nets, CNN-LSTM architectures ...)

Conclusion

Questions ?

Quantification of devastating climate events under climate change through novel multivariate bias correction methods

Séminaire Géolearning

13/11/2023, MinesParis

Grégoire Jacquemin

Supervisors : Mathieu Vrac, Denis Allard & Xavier Freulon
Hosting laboratories : Centre de géosciences, Mines Paris & LSCE
Financing organism: Mines Paris, via la Chaire Geolearning INRAE / Mines Paris

GEOLEARNING
CHAIRE /// Data Science for the Environment

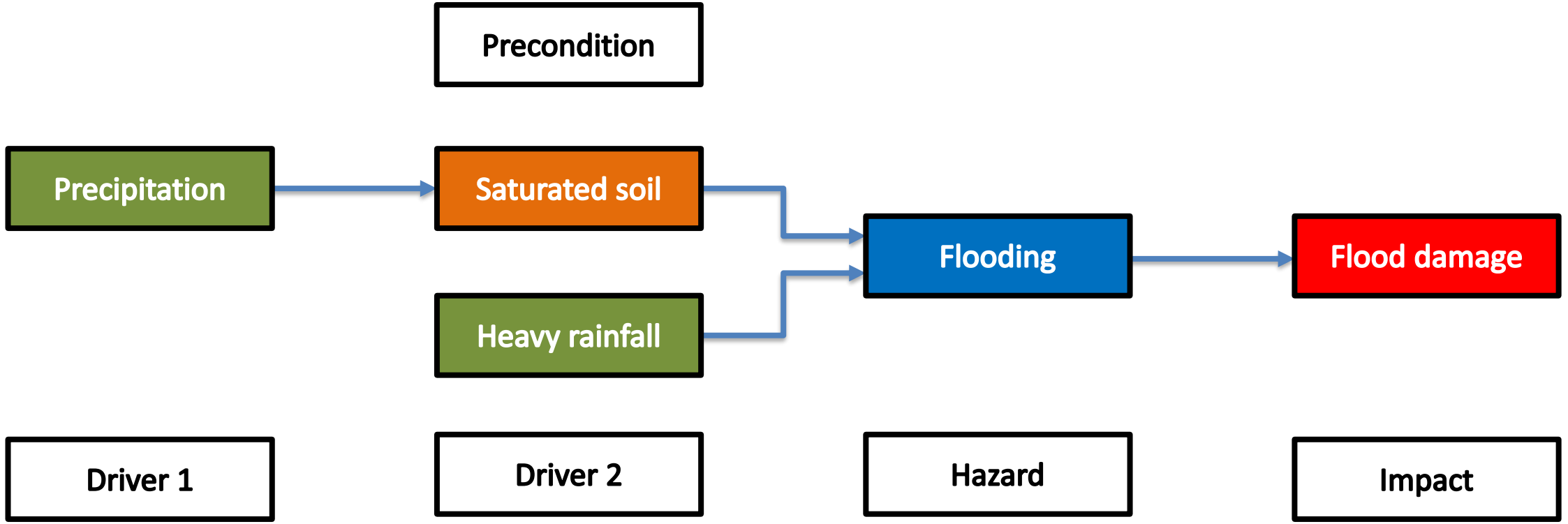
LSCE

INRAE

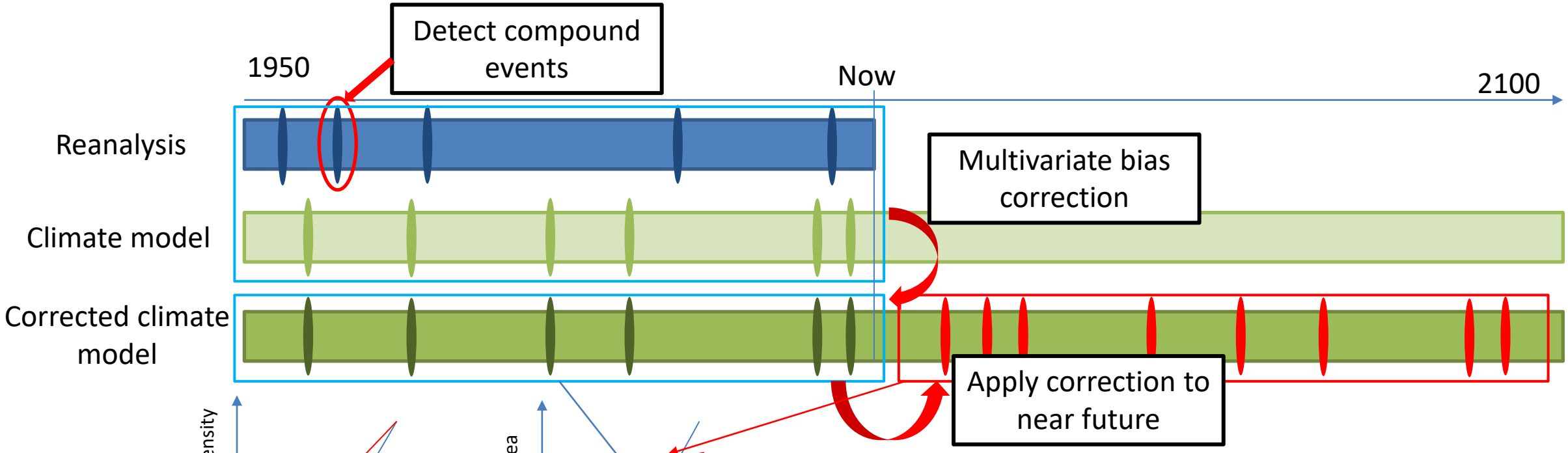
MINES PARIS | PSL

Compound events

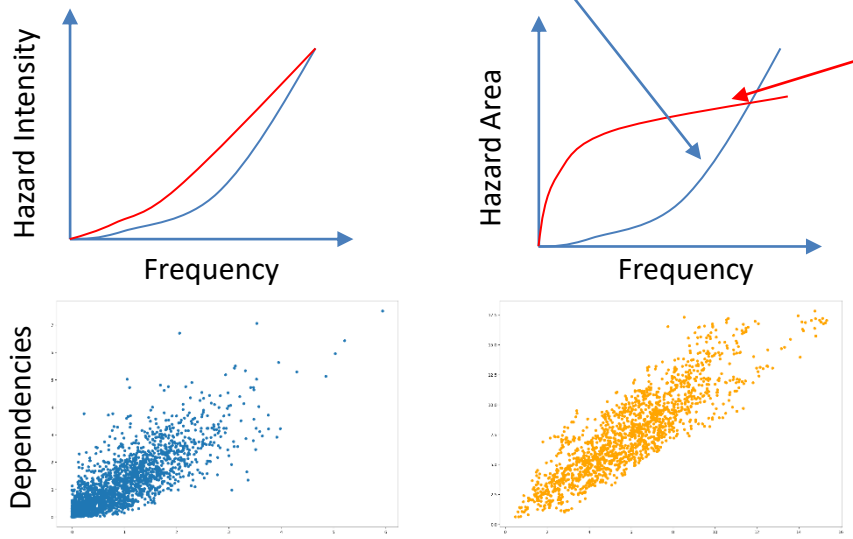
A combination of multiple drivers and/or hazards that contributes to societal or environmental risk (Zscheischler et al., 2020)

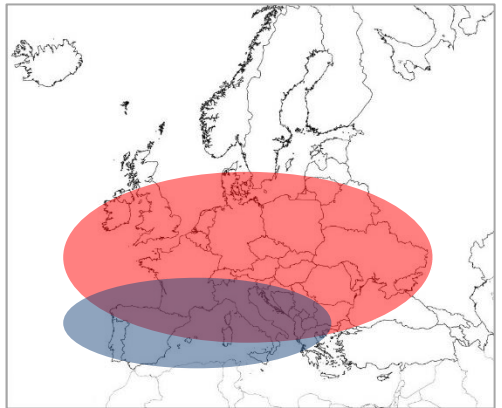


PHD Objective



Statistical analysis





Near past

Near future

Modelling the dependence

- With climate change, or in the simulations, the marginals and the dependence structure can change.
- Multivariate bias correction is probably necessary to affect both the marginals and the dependence.

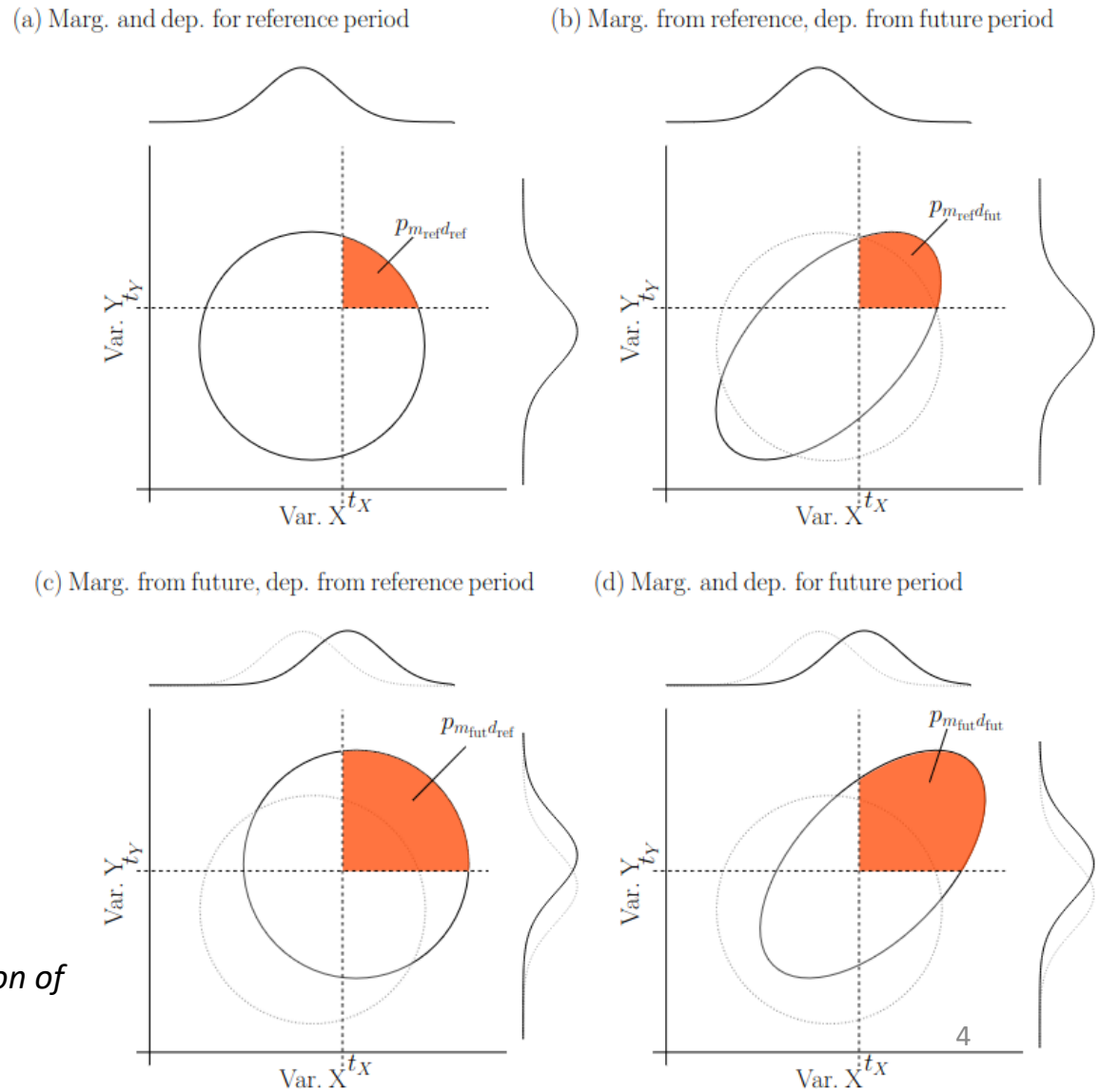


Figure from B. François and M. Vrac, *Time of emergence of compound events: contribution of univariate and dependence properties*

Multivariate extreme value theory

- X_N and Y_N are two samples with bivariate cumulative distribution function **F**, supposedly in the domain of attraction of a bivariate extreme value cumulative distribution function **G**.

- Sklar's theorem (1959) : Any multivariate cumulative distribution function **F** can be expressed in terms of its margins F_i and a copula **C** : $\boldsymbol{F}(x_1, \dots, x_d) = \boldsymbol{C}(F_1(x_1), \dots, F_d(x_d))$ with **C** : $[0,1]^{d} \rightarrow [0,1]$

- Therefore, we get weak convergence of the marginal distribution functions and the copula function

Copulas and uniform margins

- This allows us to propose the following approach :

1. Propose a univariate extreme model for the marginals
2. Reduce to uniform margins
3. Determine the copula

- X_N and Y_N must be i.i.d. and extreme, and (X_N, Y_N) must be extreme in some sense

Three potential events

- July 2021 Belgian/German flooding

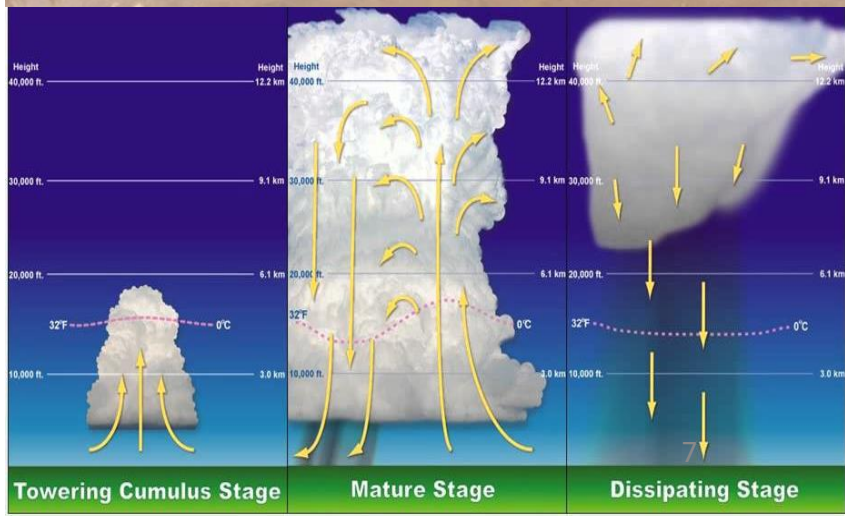
(Preconditioned event)

- May/June 2016 French flooding

(Spatially compound event)

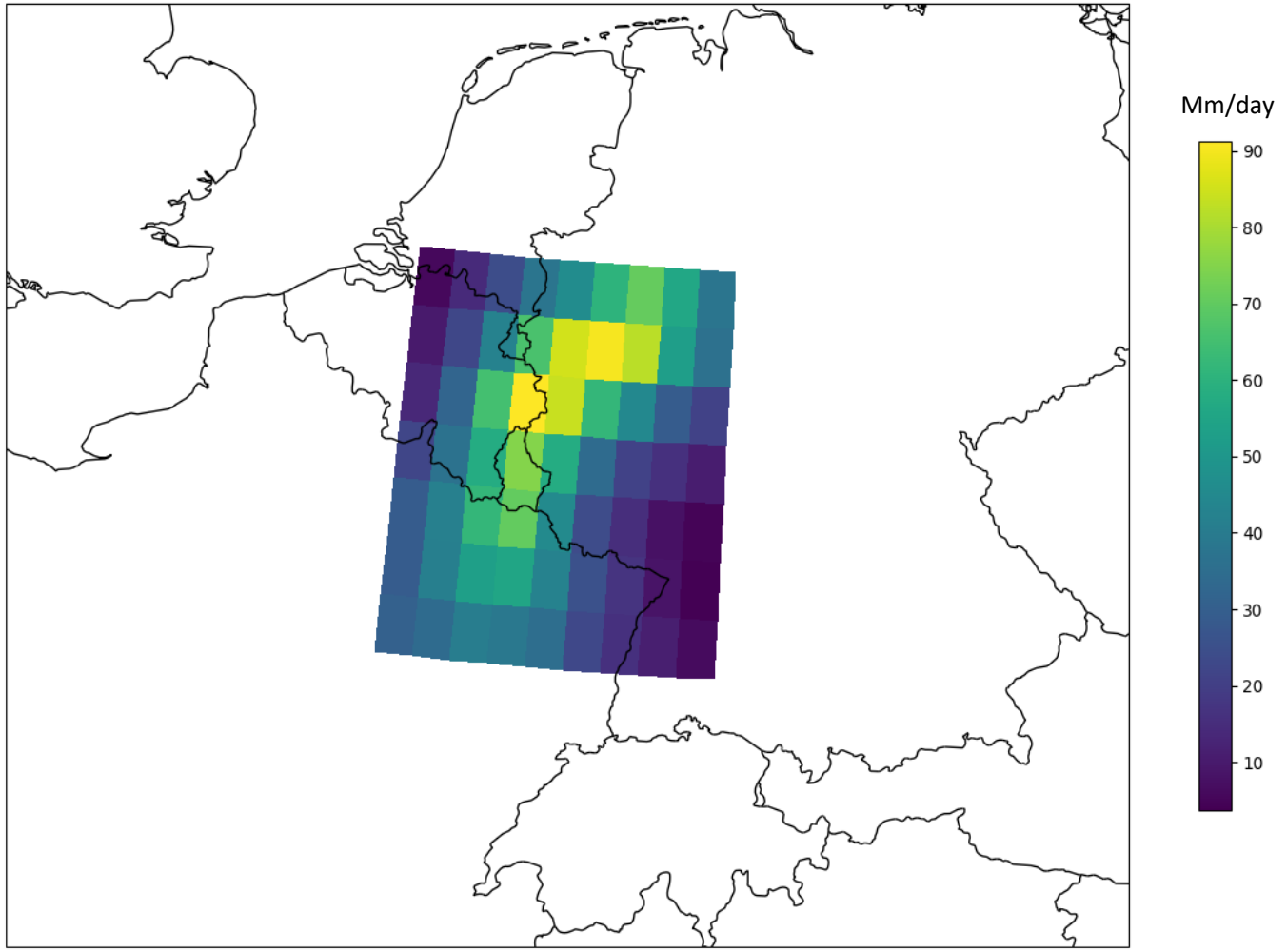
- Convective cells

(Multivariate compound event)



14th July 2021 flooding

- Data from ERA5 reanalyzes
- Total Precipitation (TP) : daily precipitation (mm/day)
- Antecedent Precipitation Index (API) : $API_j = \sum_{i=1}^{i=N} k^{i-1} * TP_{j-i}$ with k=0.9 and N=30



Data Selection

Data must be i.i.d. and extreme

Data selection for 1D

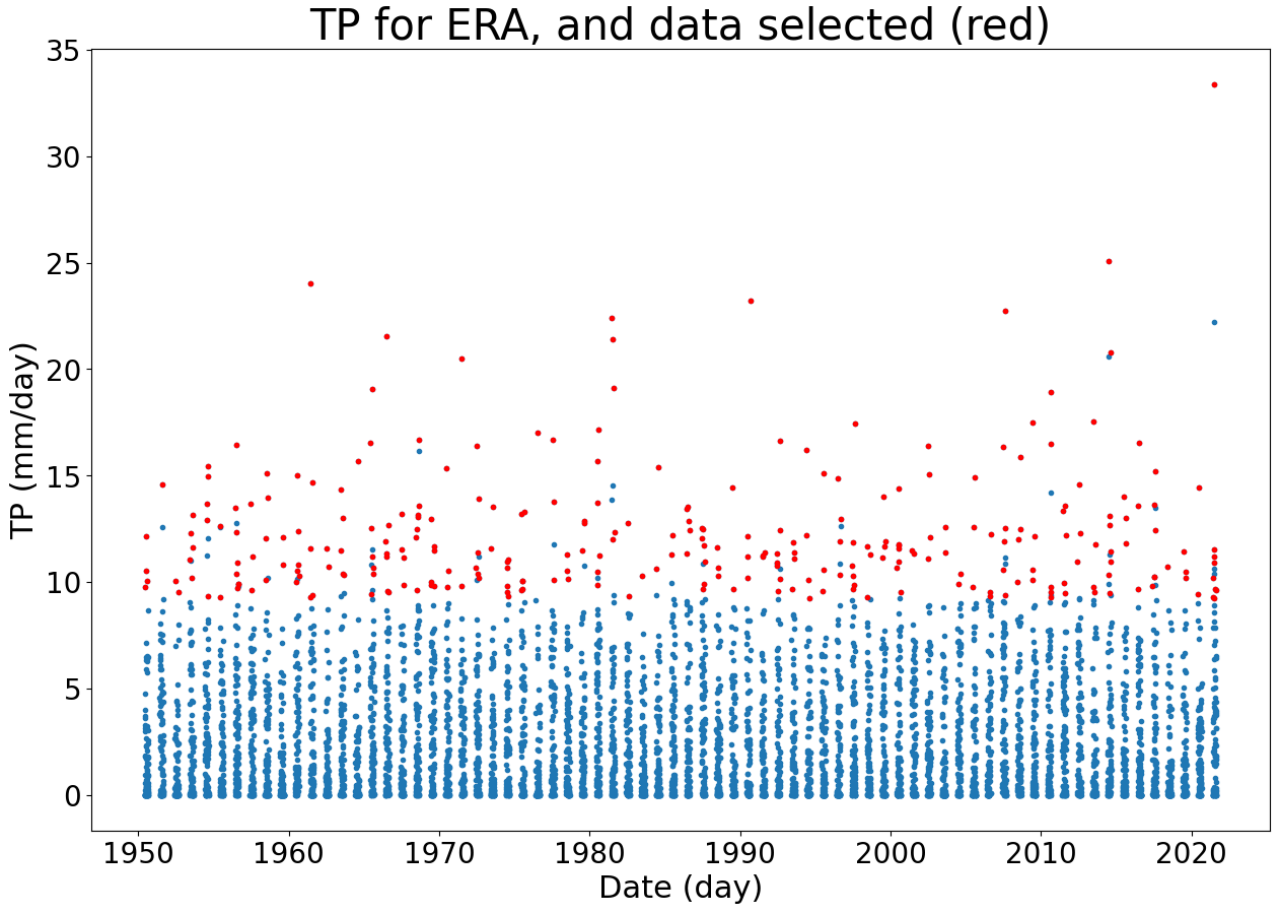
- For TP : select points above the 95th quantile, separated by at least 2 days
- For API : select points above the 95th quantile, separated by at least D days, with :

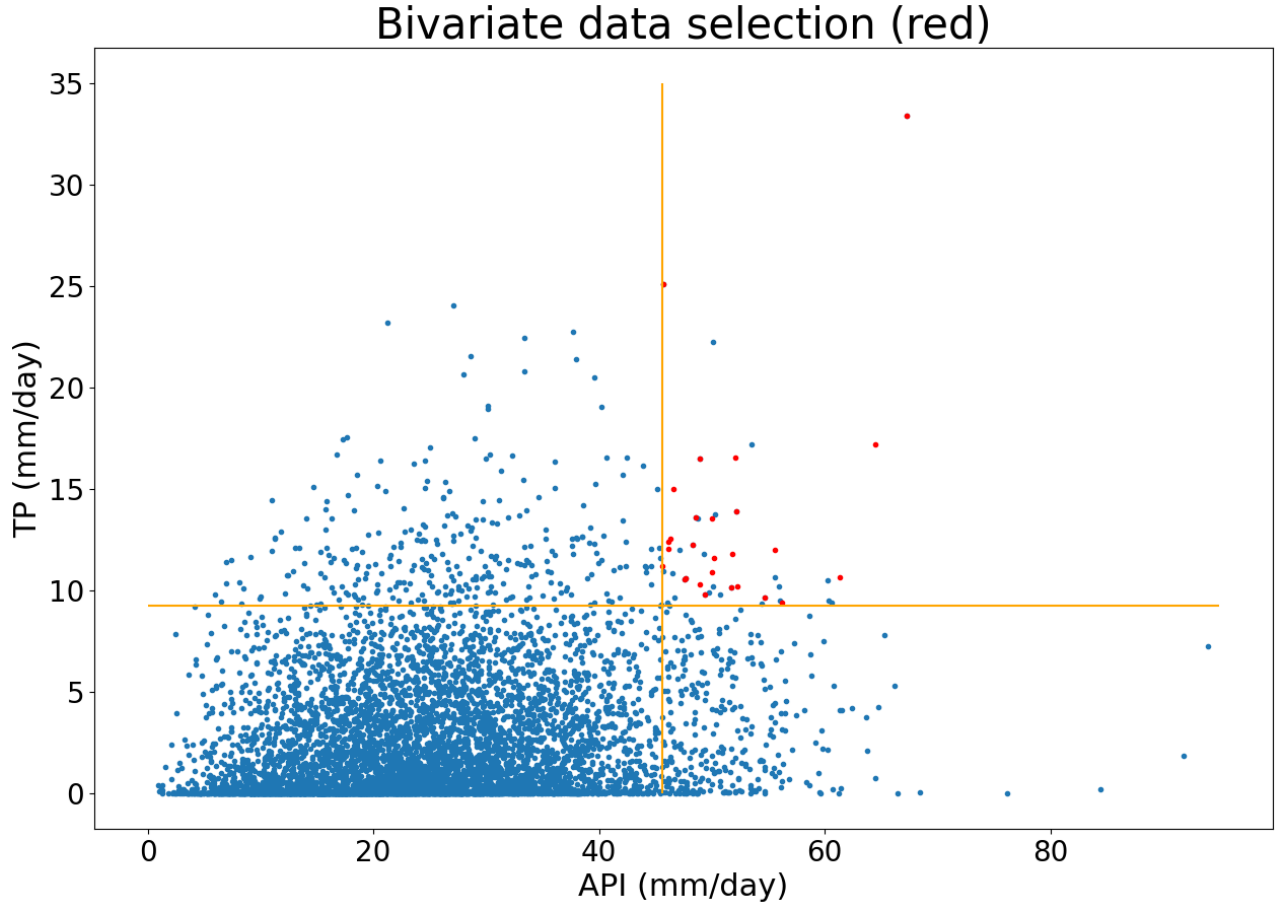
$$\rho\left(API_j, API_{j+D}\right) < 0{,}1 \quad \text{(D = 20)}$$

Data selection for 2D

- Select (TP_i, API_j) with $TP_i > Q95_{TP}, API_j > Q95_{API}$ and $i - 5 \leq j \leq i$
- Then select couples separated by at least D days, according to the highest TP value

Data selection





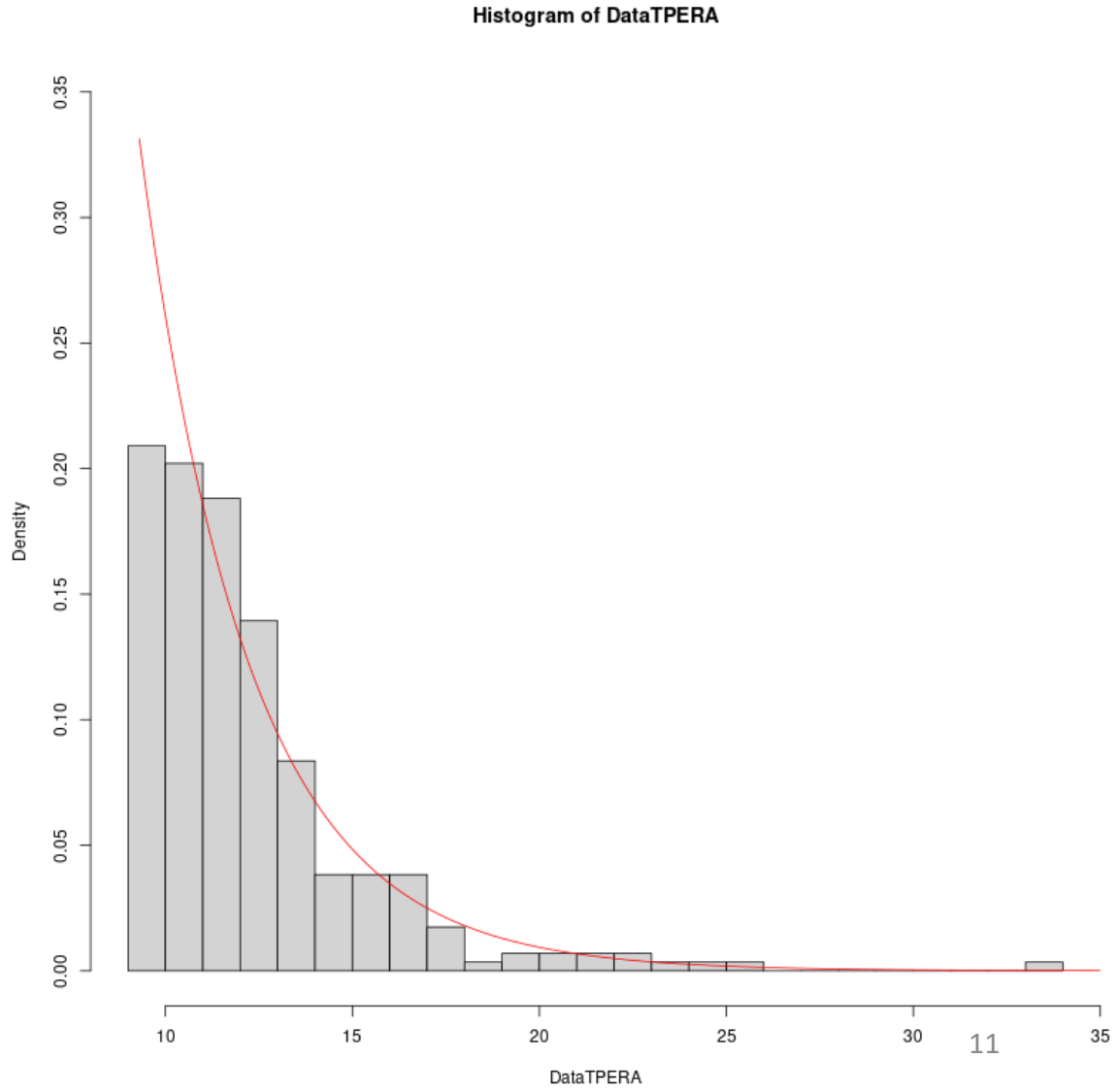
Generalized Pareto Distribution model

With the univariate data selection, we can use a Generalized Pareto Distribution (GPD) model :

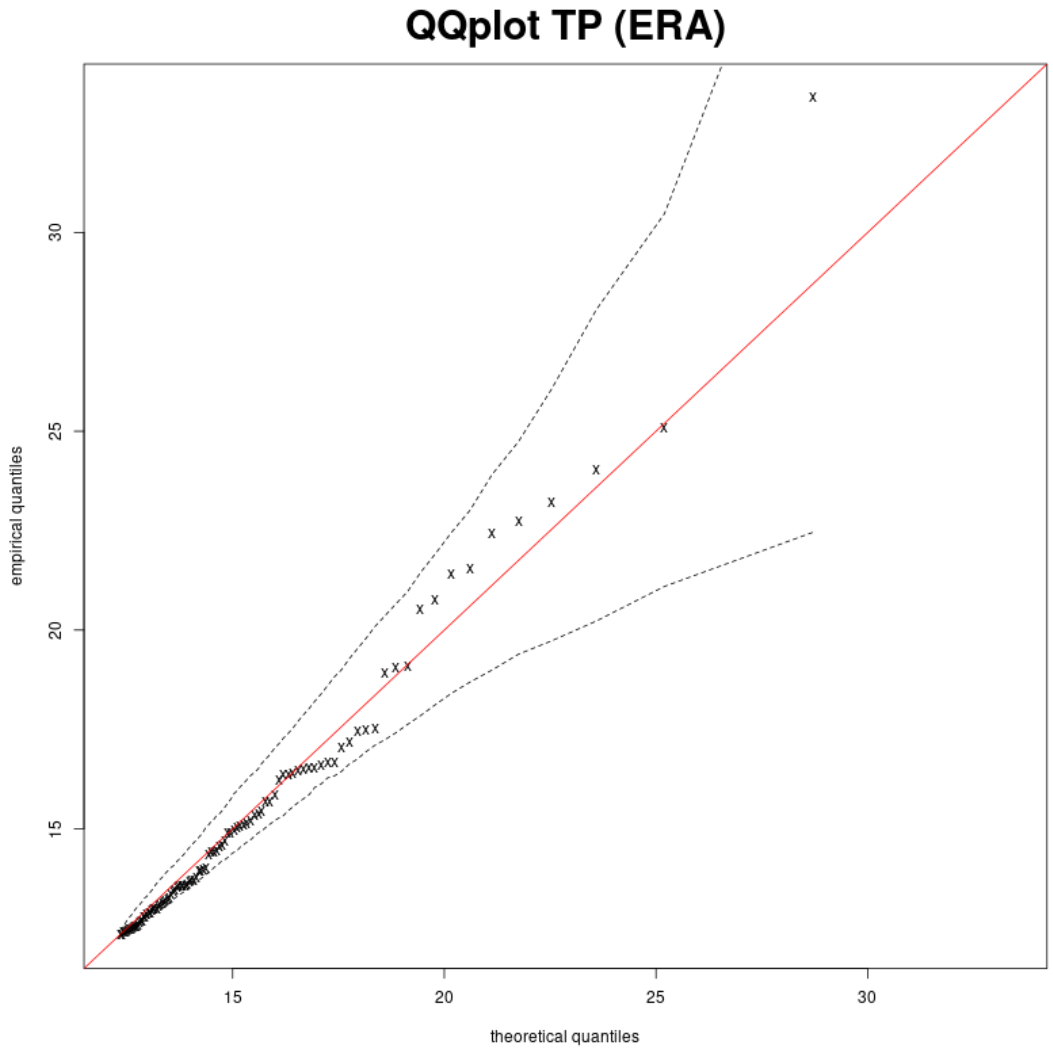
$$F(x) = 1 - (1 + \xi x)^{\frac{-1}{\xi}}$$

with $x \geq 0$ and $\xi \neq 0$

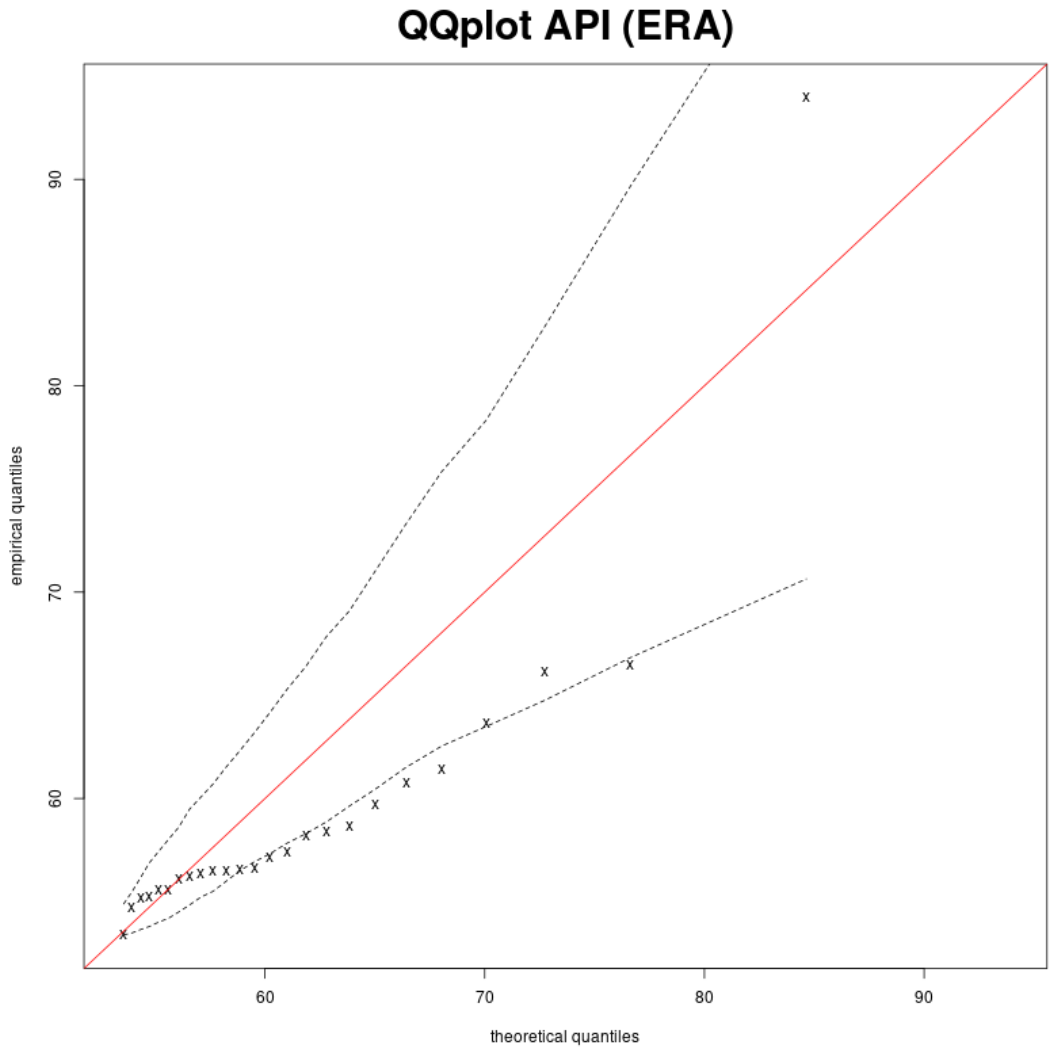
Parameters are estimated through maximum likelihood method



Quantile plots of GPD adjustment



$\hat{\xi} = 0{,}016\ (0{,}060)$

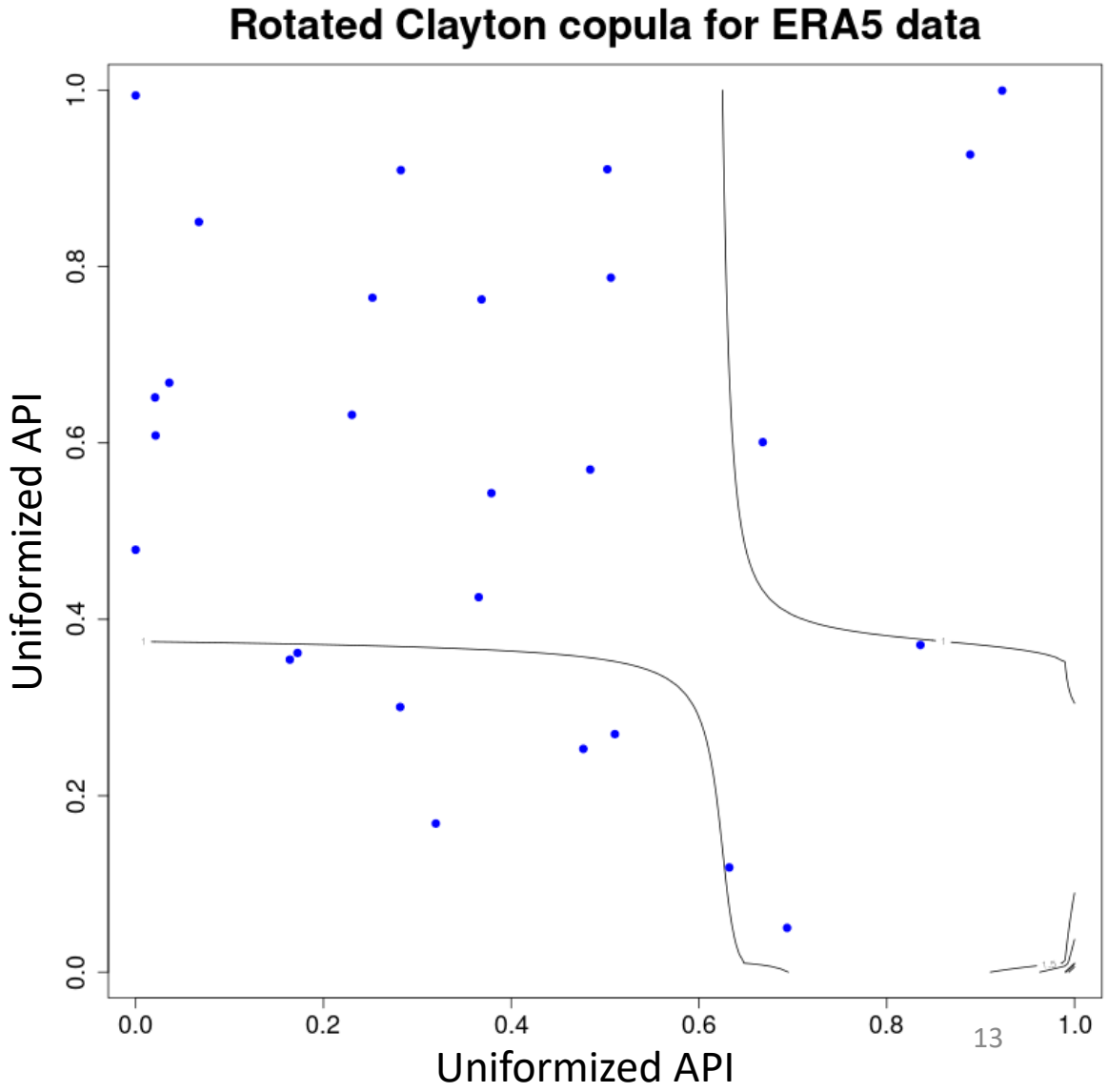


$\hat{\xi} = -0{,}063\ (0{,}130)$

Copula model estimation

Use maximum likelihood to estimate the parameters of all the copulas from the selection : Gaussian, student, Archimedean

Then select the best copula according to the Akaike Information Criteria (AIC)



Coefficients of extremal dependence

For (U, V) uniform r.v., we define χ : $\chi = \lim_{u \to 1} P(V > u | U > u)$

and $P(V > u | U > u) \approx 2 - \frac{\log C(u,u)}{\log u}$

Similarly, we can define $\bar{\chi}$: $\bar{\chi} = \lim_{u \to 1} \frac{2\log(1-u)}{\log \bar{C}(u,u)} - 1$

With $\bar{C}(u, v) = 1 - u - v + C(u, v)$

Here, we have : (χ, $\bar{\chi}$) = (0, -0.019) ---> asymptotic independence, close to total independence

Return periods

- Univariate return period = inverse of the probability to exceed a determined threshold :

$$T(x_{14.07}) = \frac{1/n}{1 - P(X \leq x_{14.07})}$$

- When describing a bivariate event by a joint exceedance (AND), the return period is defined by :

$$T_B(TP_{14.07}, API_{14.07}) = \frac{1/n}{1 - U_{TP} - U_{API} + C(U_{TP}, U_{API})}$$

with $U_X = F(x_{14.07})$ and C the copula

CDF-t correction

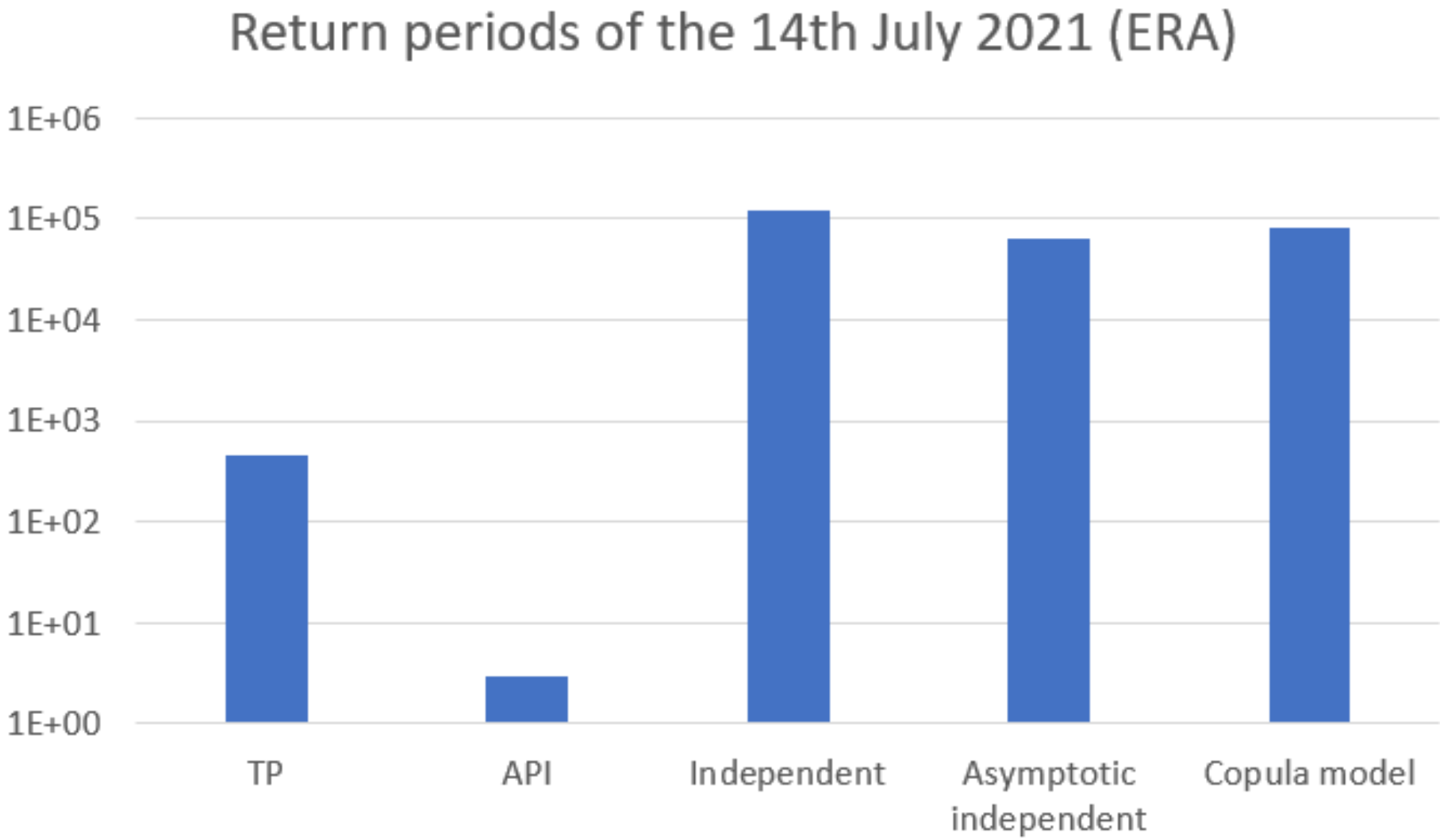
- We apply the same statistical treatment to CMIP-6 Historic data (1950-2021) and CMIP-6 Projection data (2022-2100)
- For the moment, we have considered only the IPSL model, low resolution (ssp585)

CDF-t	Historic	Projection
Model (CMIP-6)	$F_{CMIPHist}$ T	$F_{CMIPProj}$ T
Reference (ERA5)	F_{ERA}	$F_{Corrected}$

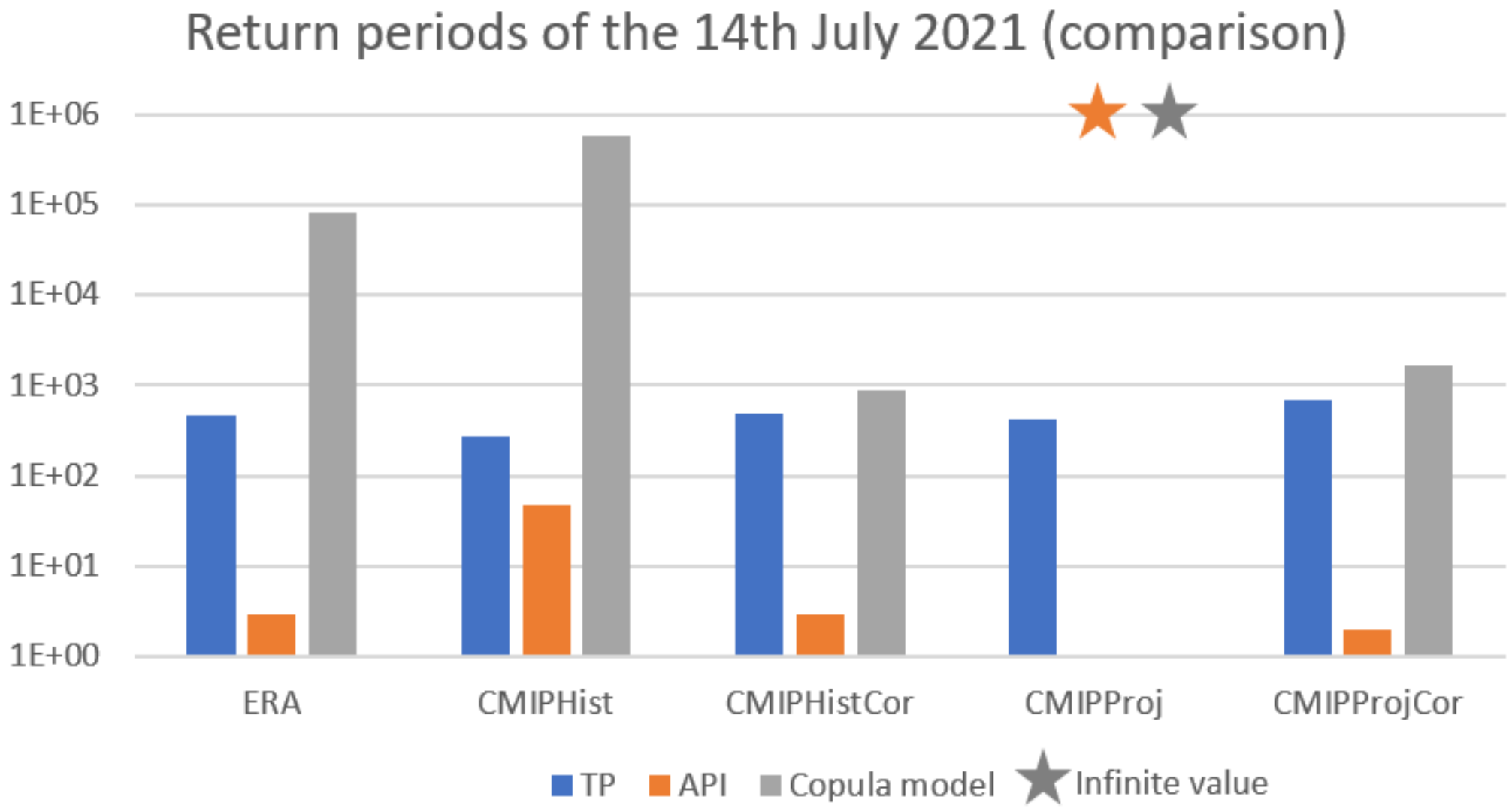
$$F_{Corrected}(x) = F_{ERA}(F^{-1}{}_{CMIPHist}\left(F_{CMIPProj}(x)\right))$$

- We get the corrected CDF, and then we perform a quantile-quantile correction between the corrected CDF and the projection data

Return periods results



Return periods results

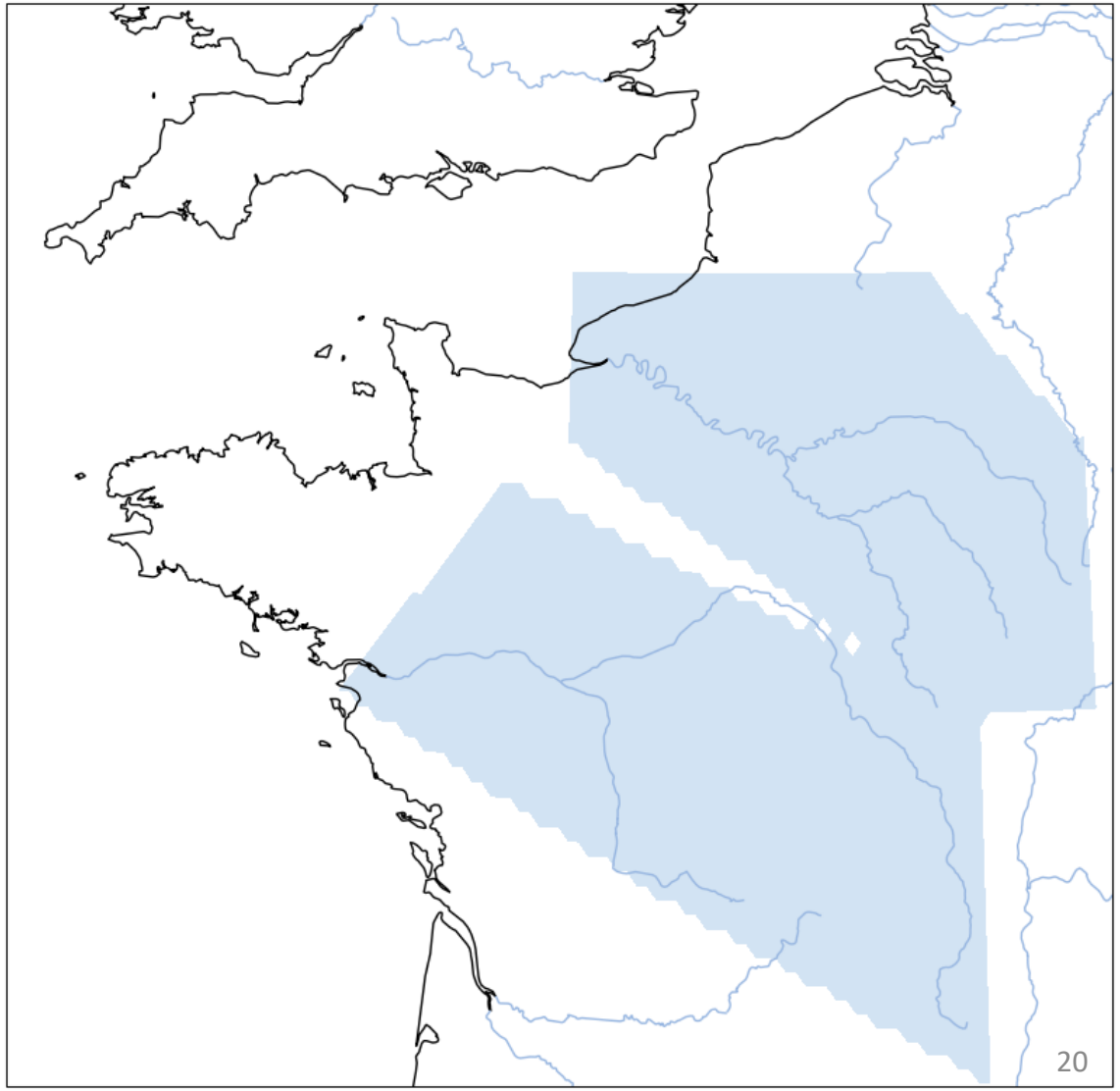


May/June French flooding

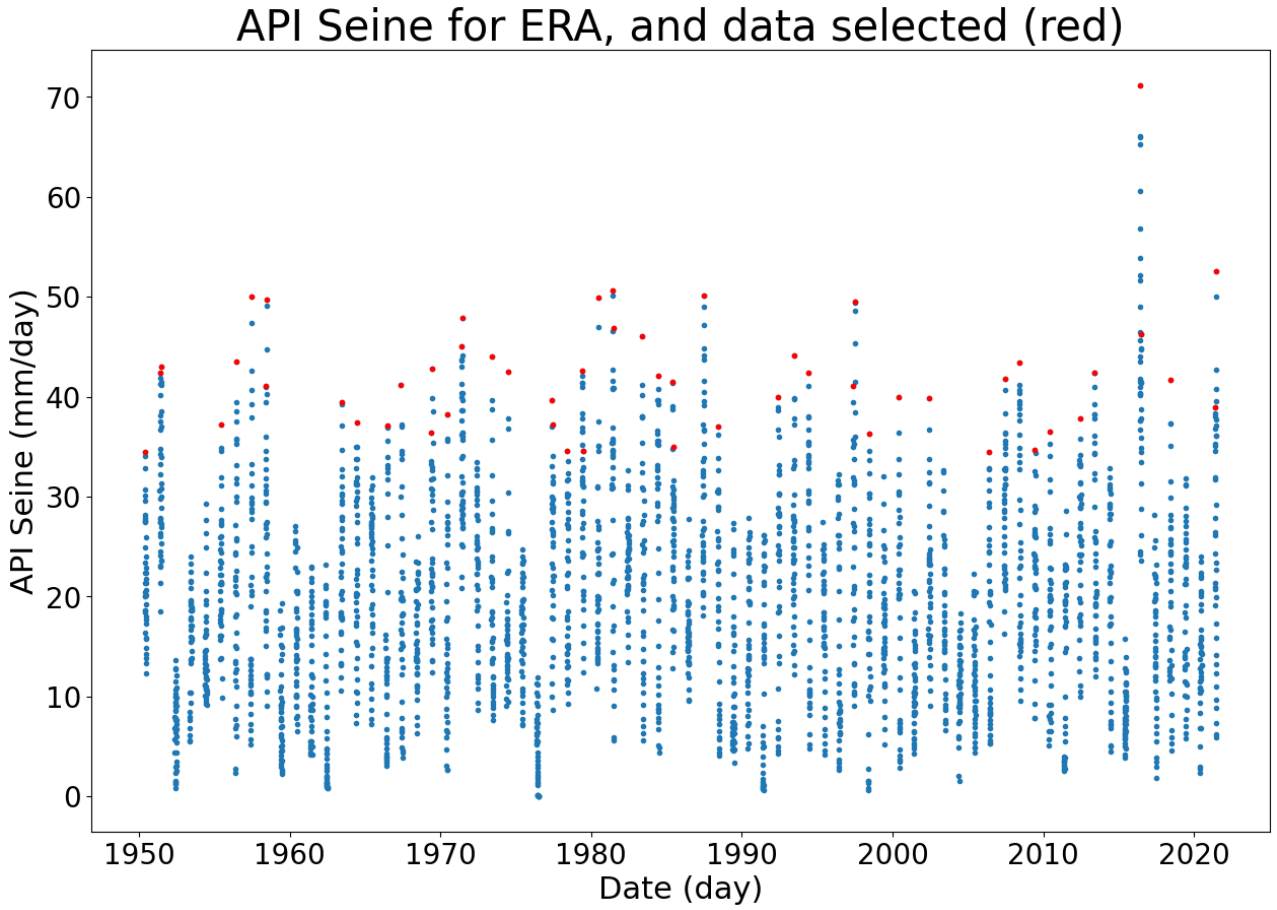
- Spatial daily precipitation means over the Seine and the Loire watersheds

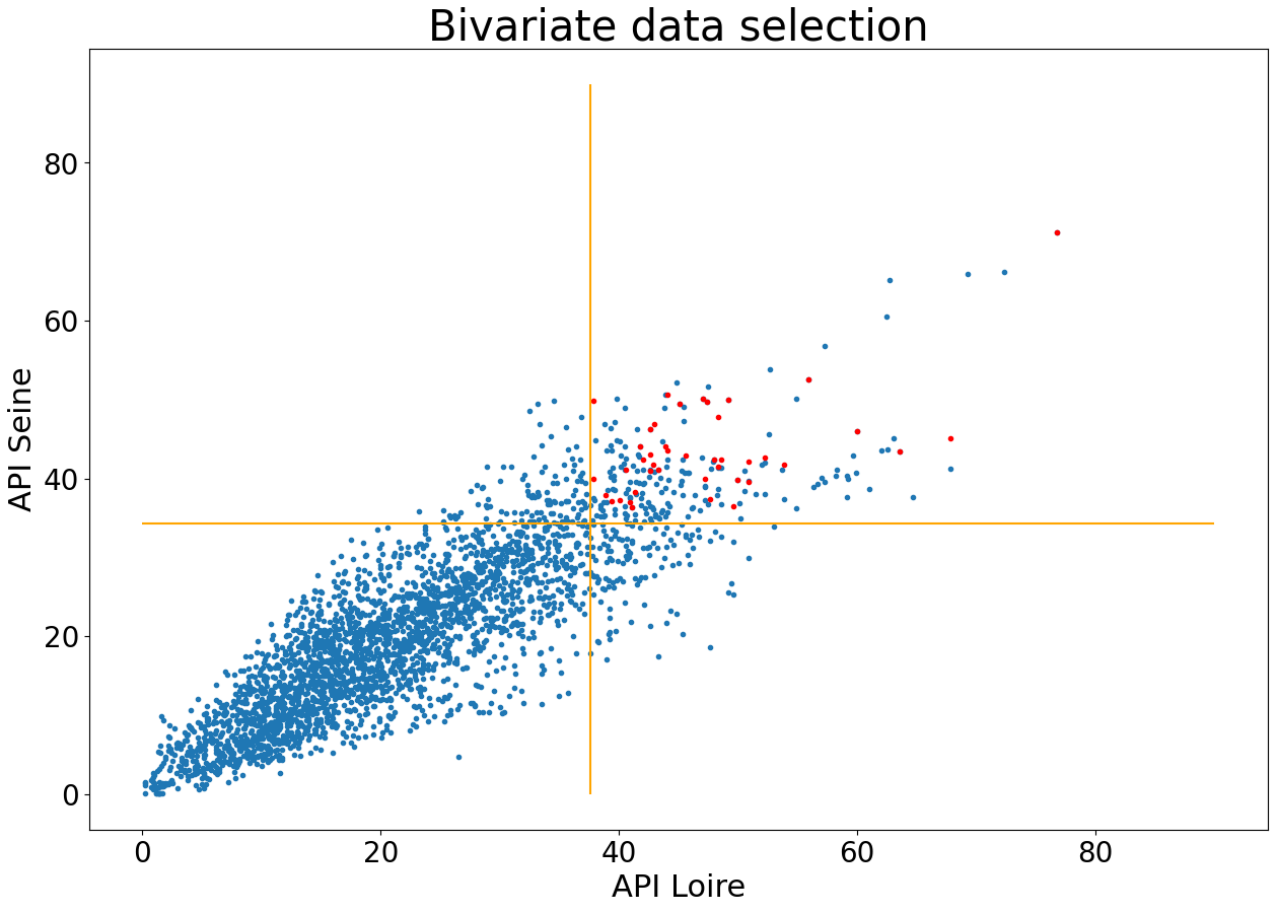
- API : $API_j = \sum_{i=1}^{i=N} k^{i-1} * TP_{j-i}$ with k=0.9 and N=20

- Same methodology (Data selection, GPD model, copula …)

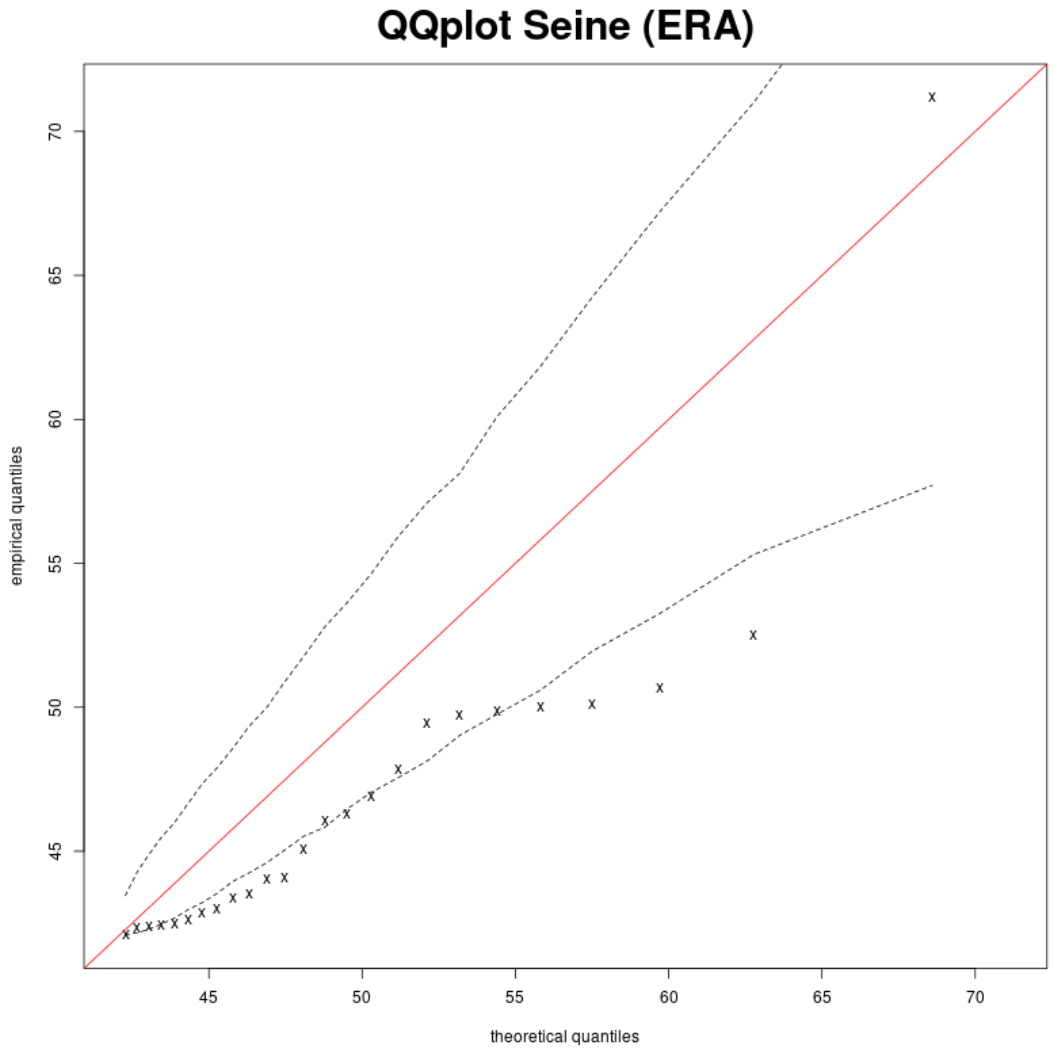


Data selection

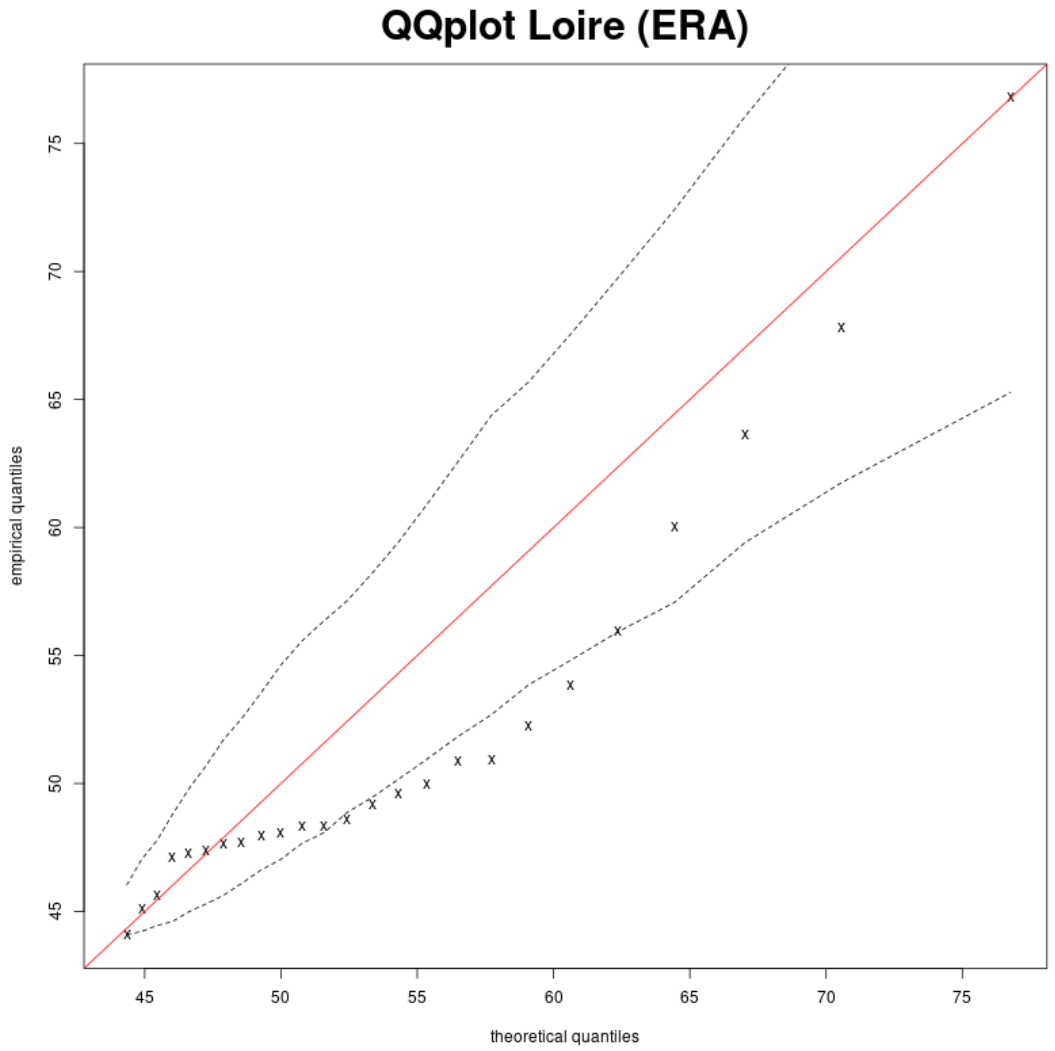




Quantile plots of GPD adjustment

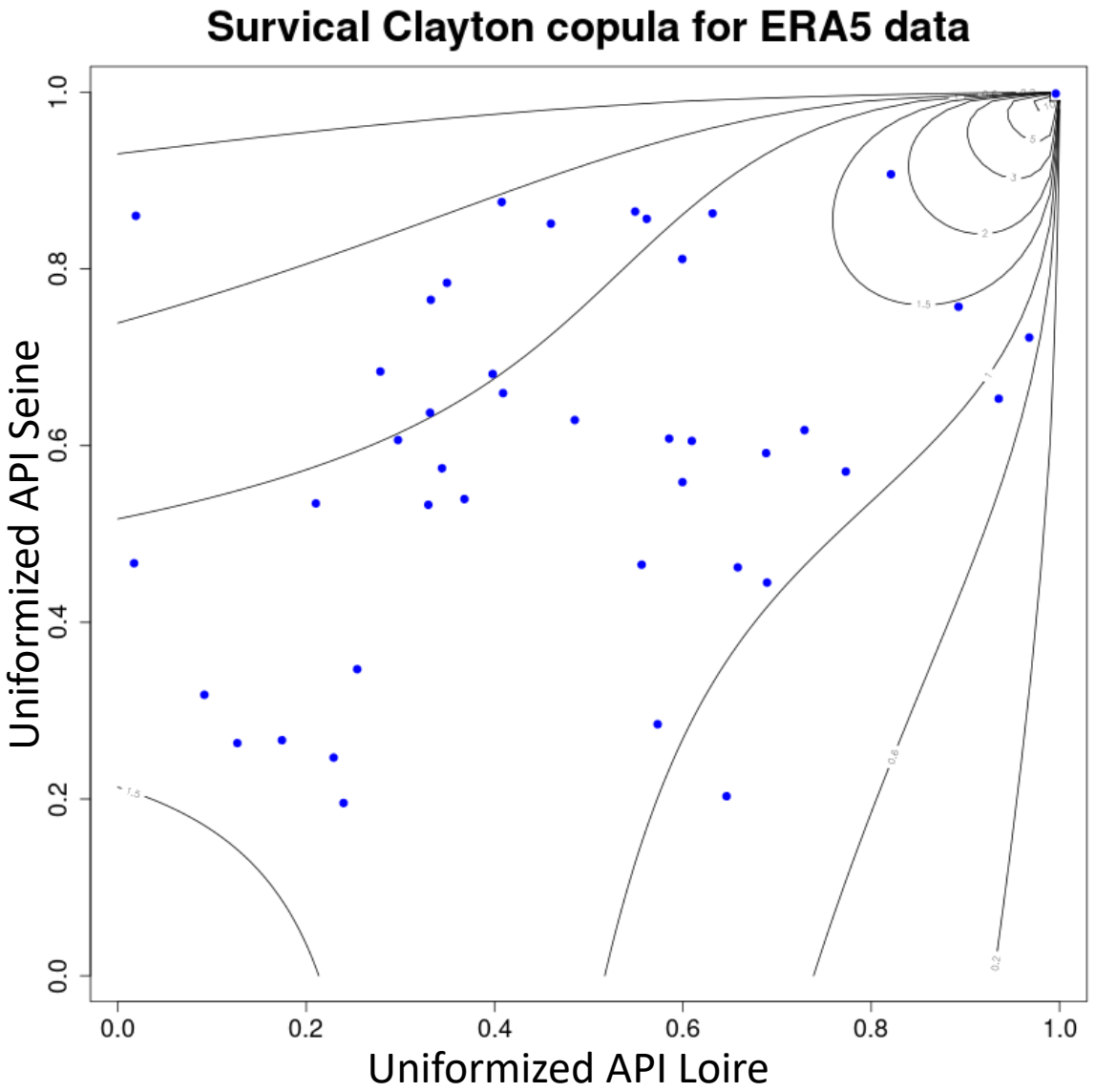


$\hat{\xi} = -0{,}164\ (0{,}116)$



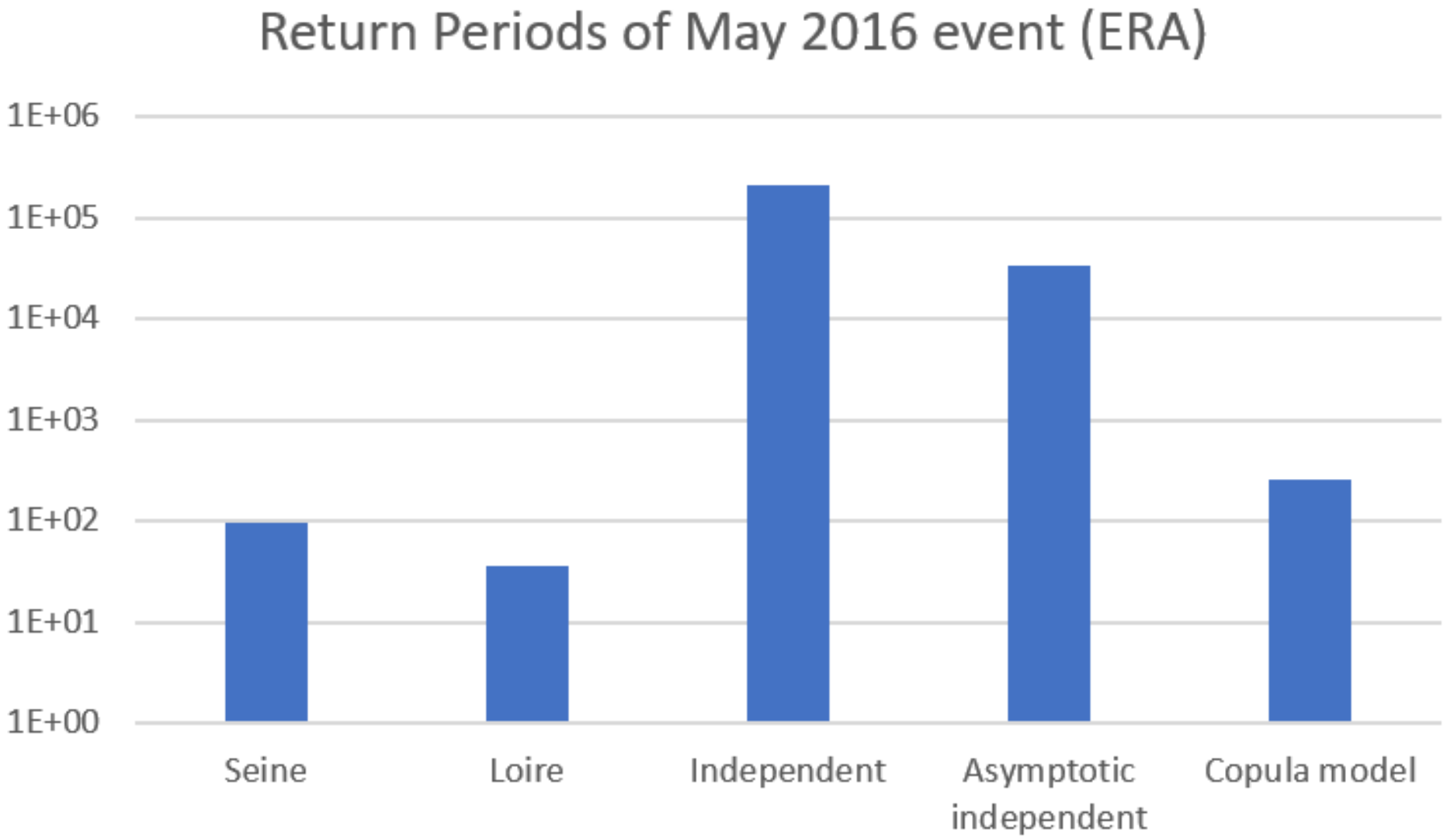
$\hat{\xi} = -0{,}252\ (0{,}109)$

Copula model

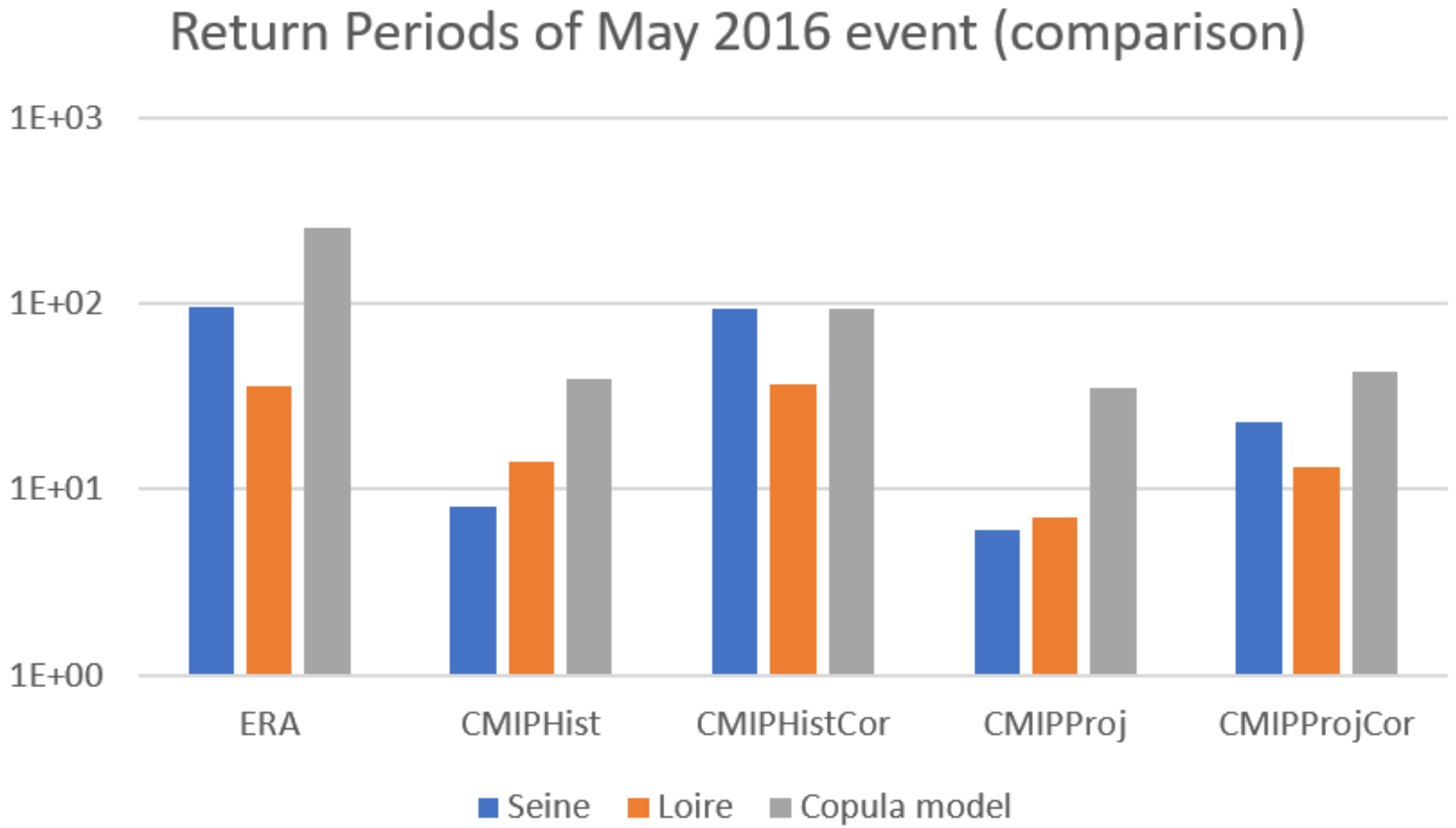


Data sets	Copula	χ	$\overline{\chi}$
ERA	Survival Clayton	0,43	1
Historic	Survival Clayton	0,295	1
Projection	Joe	0,425	1
Projection corrected	Student	0,441	1

Return periods results



Return periods results



Next steps

Next months :

- Multivariate bias correction
- Scale up framework to include more CMIP-6 simulations
- Paper

Next years :

- Apply treatment to convective cells