High-dimensional clustering of compound precipitation and wind extremes over Europe based on extremal dependence between sites

Gwladys Toulemonde[1,2]
with Alexis Boulin[2,3], Elena Di Bernardino[3] and Thomas Laloe[3]

[1] *Université Montpellier, CNRS*
[2] *INRIA, Lemon, Montpellier*
[3] *Université Côte d'Azur*

Data Science pour les risques côtiers, 2023

- Main objective : **Spatial clustering** of multivariate temporal processes
  - Space-time context
  - Compound precipitation and wind speed extremes

- based on recent development about AI-block models (Boulin et al., 2023)
- dependence summary measures appropriated for extreme value random vectors

Outline

ERA5 dataset

- We utilise the ERA5 reanalysis dataset to investigate the relationship between daily precipitation sums and daily wind speed maxima during the extended winter season (November-March).
- Available on a spatial resolution of $0.25°$ on a regular grid, and we focus on the box $[-15°E, 42.5°E] \times [30°N, 75°N]$ which covers Europe.

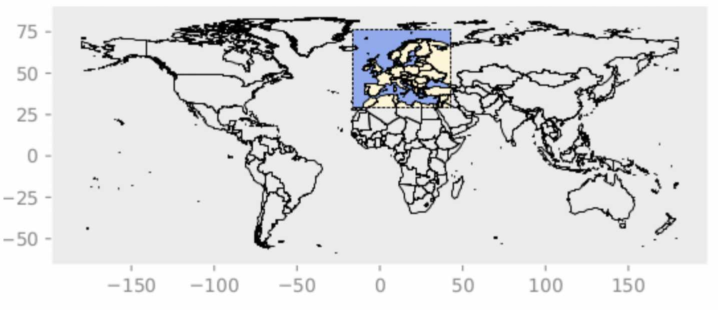


Figure 1: Considered area in the study analysis.

ERA5 dataset

- Due to computational costs, we remap the original hourly data to a regularly spaced grid with $0.5°$ spatial resolution and compute daily precipitation sums and daily wind speed maxima.
- From 1979 to 2022 (from november to march).
- The resulting dataset consists of 6655 daily sums of precipitation and wind speed maxima over 91×116 pixels with the chosen spatial resolution, hence 10556 pixels to cluster.

Compounding extremes in Europe

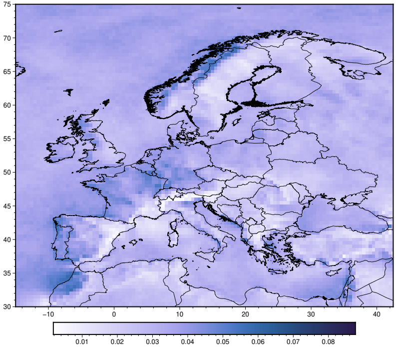
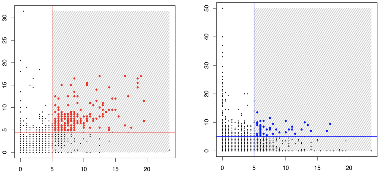


Figure 2: Proportion of both the wind speed and total precipitation that exceed their 0.9th quantiles simultaneously.

Asymptotic (in)dependence

Extremal dependence between two random variables $Y^{(1)}$ and $Y^{(2)}$. Their c.d.f are denoted by $F^{(1)}$ and $F^{(2)}$.



- The χ parameter

$$\chi = \lim_{u \to 1} \mathbb{P}\left(F^{(1)}(Y^{(1)}) > u | F^{(2)}(Y^{(2)}) > u\right)$$

$$= \lim_{u \to 1} \frac{\mathbb{P}\left(F^{(1)}(Y^{(1)}) > u, F^{(2)}(Y^{(2)}) > u\right)}{\mathbb{P}\left(F^{(2)}(Y^{(2)}) > u\right)} \equiv \lim_{u \to 1} \chi(u)$$

- $\chi > 0 \Rightarrow Y^{(1)}$ and $Y^{(2)}$ are **AD**; **the value of χ quantifies the strength of the extremal dependence.**
- $\chi = 0 \Rightarrow Y^{(1)}$ and $Y^{(2)}$ are AI.

- The extremal coefficient $\theta = 2 - \chi$

Extremal correlation between precipitation ($Z^{(j,1)}$) and wind speed ($Z^{(j,2)}$) for each site j.

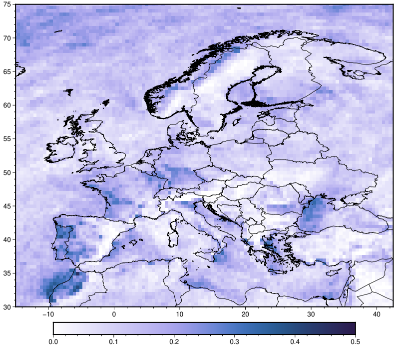


Figure 3: Estimator of the extremal correlation, $\widehat{\chi}$ between precipitation and wind. $k = 100$.

$$\widehat{\chi}(a) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{R_i^{(a,1)} > n-k+0.5, R_i^{(a,2)} > n-k+0.5\}}, \tag{1}$$

where $R_i^{(a,\ell)}$ denotes the rank of $Z_i^{(a,\ell)}$ among $Z_1^{(a,\ell)}, \ldots, Z_n^{(a,\ell)}$, $\ell = 1, 2$.

Extremal correlation according to distance between two sites

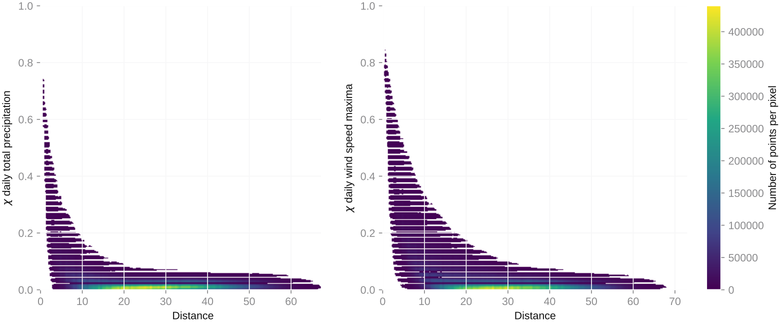


Figure 4: Estimator of the extremal correlation, $\widehat{\chi}$ for precipitation data (left) and for wind speed (right)

Dependence-based Regionalisation. Rainfall data.

- Bernard, Naveau, Vrac, Mestre, 2013
  - Extremal dependence
  - Partionning Around Medoids
- Saunders, Stephenson, Karoly, 2021
  - Extremal dependence
  - Hierarchical clustering
- Maume-Deschamps, Ribereau, Zeidan, 2023
  - Extremal concurrence probability (Dombry et al. 2018)
  - Spectral clustering
- Boulin, Di Bernardino, Laloe, Toulemonde, 2023
  - Extremal dependence
  - Presence of temporal dependence
  - Asymptotic Independent AI-block model

Max domain of attraction

- Suppose $\mathbf{Y}_n = (Y_n^{(1)}, \ldots, Y_n^{(d)})$ is a stationary multivariate random process i.d. as Y (with c.d.f. F), a d-dimensional random vector
- We assume to be in the max-domain of attraction of an EVD, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left\{ \bigvee_{i=1}^{n} \mathbf{Y}_i \leq \mathbf{u}_n(\mathbf{x}) \right\} = H(\mathbf{x})$$

  where $\mathbf{u}_n(\mathbf{x})$ a d-dimensional vector of normalising functions and H an extreme value distribution (EVD)
  - the univariate marginals $H^{(1)}, \ldots, H^{(d)}$ of H are univariate EVD
  - the dependent structure of H

$$-\ln H(x) = L(-\ln H^{(1)}(x^{(1)}), \ldots, -\ln H^{(d)}(x^{(d)}))$$

    $L : [0,\infty)^d \to [0,\infty)$ the stable tail dependence function

$$L(\mathbf{x}) = \lim_{t\to 0} t^{-1} \mathbb{P}\{F^{(1)}(Y^{(1)}) > 1 - tx^{(1)} or \ldots or F^{(d)}(Y^{(1)}) > 1 - tx^{(d)}\}$$

Asymptotic Independent block model

- $(\mathbf{Y}_n, n \in \mathbb{N})$ exhibits Asymptotic Independence (AI) when the limit distribution, the multivariate extreme value distribution H is equal to the product of its marginal EVD $H^{(1)}, \ldots, H^{(d)}$:

$$H = \Pi_{j=1}^{d} H^{(j)}$$

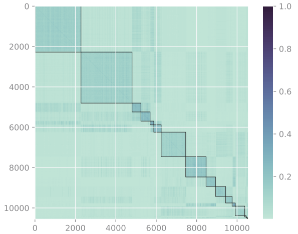
- $(\mathbf{Y}_n, n \in \mathbb{N})$ is said to follow an AI block model with G groups if there exists a partition $O = \{O_g\}_{g=1}^{G}$ of $\{1, \ldots, d\}$ with $|O_g| = d_g$ and marginal extreme value distributions $H^{(O_g)} : \mathbb{R}^{d_g} \to [0,1]$ such that

$$H = \Pi_{g=1}^{G} H^{(O_g)}$$

.

Asymptotic Independent block model

- **Method:** variable clustering in order to separate groups which can be assumed to be independent in the extremes
- **Application:** spatial clustering based on temporal processes
- **Fundamental object:** matrix of extremal correlation coefficients χ between each pair of sites
- **Proposal:** algorithm which retrieves the thinnest partition with high probability



Outline

Introduction

A measure for evaluating dependence between compound extremes

Clustering algorithm for compound extreme events

Detecting concomitant extremes of compound precipitation and wind speed extremes

Conclusions

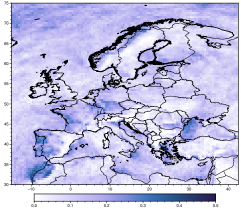
Notations

- Specifically, let $(\mathbf{Z}_n^{(s)}, s \in D \subseteq \mathbb{R}^2, n \in \mathbb{N})$ be a spatio-temporal random field.
- $\mathbf{Z}_n^{(s)} = (Z_n^{(s,1)}, Z_n^{(s,2)})$ is a vector corresponding to the daily sums of precipitation and wind speed maxima at time n at location s.
- Assume that observations are available

  over d (in our dataset, $d = 10556$) spatial locations

  for each time n (total 6655)

  We have $\mathbf{Z}_n = (\mathbf{Z}_n^{(1)}, \dots, \mathbf{Z}_n^{(d)})$, where $\mathbf{Z}_n$ is a $2d$-random vector with stationary law $\mathbf{Z} = (\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(d)})$.



Sum of Extremal COefficient

- We introduce a new coefficient called Sum of Extremal COefficient (SECO).
- The purpose of this metric is to quantify any deviation from asymptotic independence of groups of variables.
- The pairwise SECO metric is defined as

$$\mathrm{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) = \theta(a) + \theta(b) - \theta(a,b).$$

where

$$\theta(j) = \lim_{q \to 0} q^{-1} \mathbb{P}\left\{ \max_{\ell=1,2} F^{(j,\ell)}(Z^{(j,\ell)}) > 1 - q \right\}, \quad j = a, b$$

$$\theta(a,b) = \lim_{q \to 0} q^{-1} \mathbb{P}\left\{ \max_{j=a,b} \max_{\ell=1,2} F^{(j,\ell)}(Z^{(j,\ell)}) > 1 - q \right\}$$

A measure of dependence

$$\mathrm{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) = \theta(a) + \theta(b) - \theta(a,b).$$

- The SECO metric is always positive and quantifies the deviation from asymptotic independence between the two groups of variables.
- Indeed, the SECO metric is equal to zero if and only if the two groups of variables are asymptotically independent random vectors.
- Furthermore, the pairwise SECO reduces to the extremal correlation

$$\mathrm{SECO}(Z^{(1)}, Z^{(2)}) = 2 - \theta(1,2) = \chi(1,2),$$

if $Z^{(1)}$ and $Z^{(2)}$ are univariate random variables.

An empirical estimator of SECO

- The empirical counterpart of the SECO is denoted as $\widehat{\mathrm{SECO}}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)})$ and is defined as:

$$\widehat{\mathrm{SECO}}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}) = \widehat{\theta}(a) + \widehat{\theta}(b) - \widehat{\theta}(a,b),$$

where $\widehat{\theta}$ is a nonparametric estimator of the extremal coefficient θ (see for instance [Einmahl et al., 2012]) where

$$\widehat{\theta}(j) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}_{\{R_i^{(j,1)} > n-k+0.5 \text{ or } R_i^{(j,2)} > n-k+0.5\}}, \quad j = a, b, \qquad (2)$$

with $R_i^{(j,\ell)}$ the rank of $Z_i^{(j,\ell)}$ among $Z_1^{(j,\ell)}, \ldots, Z_n^{(j,\ell)}$, $j = a, b$, $\ell = 1, 2$.

- Under mixing conditions, we can show that this statistic is consistent i.e.

$$\widehat{\mathrm{SECO}}(\mathsf{Z}^{(a)}, \mathsf{Z}^{(b)}) \xrightarrow[n\to\infty]{\mathbb{P}} \mathrm{SECO}(\mathsf{Z}^{(a)}, \mathsf{Z}^{(b)}).$$

- Furthermore, we have

$$\widehat{\Theta}(a,b) := \widehat{\mathrm{SECO}}(\mathsf{Z}^{(a)}, \mathsf{Z}^{(b)}) / \min\{\widehat{\theta}(a), \widehat{\theta}(b)\} \in [0,1].$$

Pairwise SECO between sites

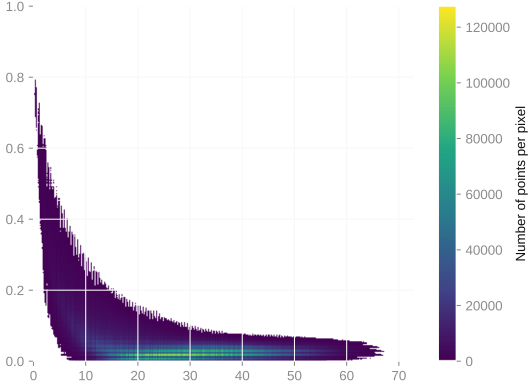


Figure 5: Pairwise empirical SECO using the 100 greatest value with respect to the pairwise distance.

Outline

Introduction

A measure for evaluating dependence between compound extremes

Clustering algorithm for compound extreme events

Detecting concomitant extremes of compound precipitation and wind speed extremes

Conclusions

Constrained Asymptotic Independent block model

A constrained asymptotic independent block model is defined by

- $\mathbf{Z}$ a $2d$-vector with law F having d marginal random vectors (d sites): $\mathbf{Z} = (Z^{(1,1)}, Z^{(1,2)}, Z^{(2,1)}, Z^{(2,2)}, \dots, Z^{(d,1)}, Z^{(d,2)})$
- F is in the max domain of attraction of H.
- There exists $O = \{O_g\}_{g=1}^{G}$ a partition of $\{1, \dots, d\}$ in G groups with $|O_g| = d_g$ and marginal extreme value distributions $H^{(O_g)} : \mathbb{R}^{d_g} \to [0,1]$ such that $H = \Pi_{g=1}^{G} H^{(O_g)}$.

Bivariate comparison to retrieve the hidden partition

In a constrained asymptotic independent block model, the following statement

$$\mathsf{Z}^{(O_{g_1})} \perp\!\!\!\perp_{ext} \mathsf{Z}^{(O_{g_2})}$$

is equivalent to the following statement, implying the SECO

$$\mathrm{SECO}(a,b) = \mathrm{SECO}(b,a) = 0, \forall a \in O_{g_1}, \forall b \in O_{g_2}.$$

Thus, the SECO is a sufficient metric to derive a simple, yet powerful, algorithm to recover the hidden partition.

The algorithm

- Algorithm CAICE (Clustering procedure for AI block models with compound extremes) for $S = \{1, \ldots, d\}$
  - Based on the normalised SECO

$$\widehat{\Theta}(a,b) = \widehat{\mathrm{SECO}}(a,b) / \min\{\widehat{\theta}(a), \widehat{\theta}(b)\}, \quad a,b \in \{1,\ldots,d\}, \tag{3}$$

  - No choice for the number of groups
  - Involving a threshod τ

CAICE(S, τ, $\widehat{\Theta}$)

Algorithm (CAICE) Clustering procedure for AI block models with compound extremes

procedure CAICE(S, τ, $\hat{\Theta}$)
    Initialise: $S = \{1, \ldots, d\}$, $\hat{\Theta}(a, b)$ for $a, b \in \{1, \ldots, d\}$ and $l = 0$
    while $S \neq \emptyset$ **do**
        $l = l + 1$
        if $|S| > 1$ **then**
            $(a_l, b_l) = \arg\max_{a,b \in S} \hat{\Theta}(a, b)$
            if $\hat{\Theta}(a_l, b_l) > \tau$ **then**
                $\hat{O}_l = \{s \in S : \hat{\Theta}(a_l, s) \geq \tau \text{ and } \hat{\Theta}(b_l, s) \geq \tau\}$
            if $\hat{\Theta}(a_l, b_l) \leq \tau$ **then**
                $\hat{O}_l = \{a_l\}$
        if $|S| = 1$ **then**
            $\hat{O}_l = S$
        $S = S \setminus \hat{O}_l$
    return $\hat{O} = (\hat{O}_l)_l$

How to set up the threshold τ?

- Set $\tau > 0$ the threshold parameter in the algorithm.
- For this threshold τ, the algorithm returns a partition $\widehat{O}_1, \ldots, \widehat{O}_G$ of $\{1, \ldots, d\}$ with respective sizes d_g.
- With this partition, in each cluster g, $\mathbf{X}_\tau^{(g)}$ is a $(2d_g)$-dimensional random vector for which we can compute $\widehat{\theta}(g)$.
- The empirical SECO of this partition is

$$\widehat{\mathrm{SECO}}(\mathbf{X}_\tau^{(1)}, \ldots, \mathbf{X}_\tau^{(G)}) = \sum_{g=1}^{G} \widehat{\theta}(g) - \widehat{\theta}(1, \ldots, d).$$

- Find τ which minimizes the empirical SECO permit to recover an AI-block model

Outline

Calibration of the threshold τ

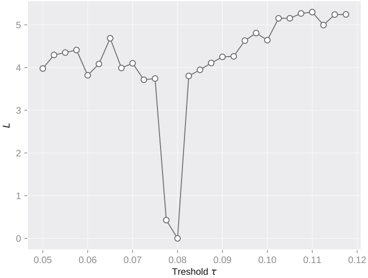


Figure 6: Value of the function L for $\tau \in \Delta = \{0.05, 0.0525, \ldots, 0.12\}$ for the 30 greatest values.

with

$$L(\tau) = \ln\left(1 + \left(\widehat{\mathrm{SECO}}(\mathbf{X}_{\tau}^{(1)}, \ldots, \mathbf{X}_{\tau}^{(G)}) - \min_{\tau \in \Delta} \widehat{\mathrm{SECO}}(\mathbf{X}_{\tau}^{(1)}, \ldots, \mathbf{X}_{\tau}^{(G)})\right)\right).$$

Clustered pairwise SECO

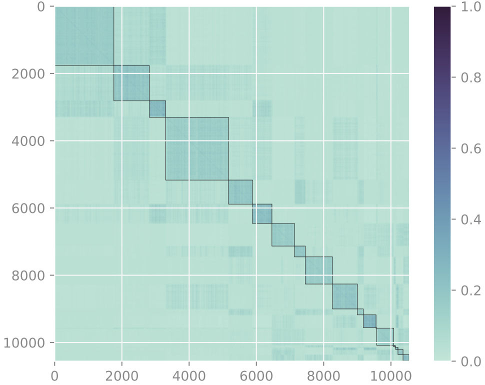


Figure 7: Partition of the SECO similarity matrix with threshold $\tau = 0.08$. Squares represent the clusters of variables.

Spatial representation of clusters $\widehat{O}^{PW}$

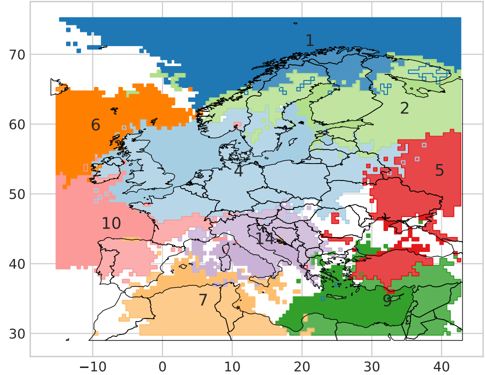


Figure 8: Representation of the 9 largest clusters (in decreasing order) of the partition of the SECO matrix between daily precipitation sums and wind speed maxima with threshold $\tau = 0.08$.

Hierarchical clustering on the fourth cluster

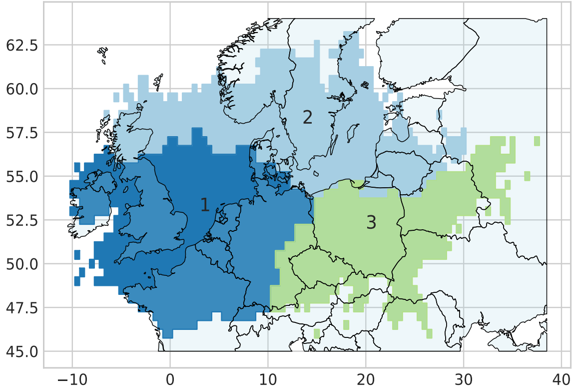


Figure 9: Representation of the 3 clusters of the partition of the 1868 pixels of the fourth cluster of the partition given by Algorithm CAICE using extremes of daily total precipitation and wind speed maxima. $1 - \widehat{SECO}$ is used as the dissimilarity matrix

Outline

What has been presented

- Proposition of the SECO
- Spatial clustering of multivariate processes based on extremal dependence
- Identify areas within Europe that exhibit independence regarding the extremes of compound precipitation and wind speed.

What has not been presented

- Quantify the role of Precipitation and the role of wind in the construction of the partition using the Adjusted Rank Index (ARI), a concordance score between two different partitions.
- The natural extension to p_j marginal univariate random variables in each site.

References

- Bernard, E., P. Naveau, M. Vrac, and O. Mestre (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in France. Journal of climate 26 (20), 7929-7937
- Boulin, A., Di Bernardino, E., Laloe, T., Toulemonde, G (2023). High-dimensional variable clustering based on sub-asymptotic maxima of a weakly dependent random process. 2023. arXiv:2302.00934v2.
- Boulin, A., Di Bernardino, E., Laloe, T., Toulemonde, G (2023). Identifying concomitant compound precipitation and wind speed extremes. arXiv:2311.11292
- Maume-Deschamps, V., P. Ribereau, and M. Zeidan (2023). Detecting the station- arity of spatial dependence structure using spectral clustering. https://hal.science/hal- 03918937
- Saunders, KR, AG Stephenson, and DJ Karoly. (2021). A regionalisation approach for rainfall based on extremal dependence. Extremes 24 (2): 215-240.