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

Gwladys Toulemonde\textsuperscript{1,2} with Alexis Boulin\textsuperscript{2,3}, Elena Di Bernardino\textsuperscript{3} and Thomas Laloe\textsuperscript{3}

\textsuperscript{1}Université Montpellier, CNRS
\textsuperscript{2}INRIA, Lemon, Montpellier
\textsuperscript{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

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
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 $[-15^\circ E, 42.5^\circ E] \times [30^\circ N, 75^\circ N]$ which covers Europe.

Figure 1: Considered area in the study analysis.
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 $91 \times 116$ pixels with the chosen spatial resolution, hence 10556 pixels to cluster.
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 $\chi$ parameter

\[
\chi = \lim_{u \to 1} \frac{\mathbb{P}(F^{(1)}(Y^{(1)}) > u | F^{(2)}(Y^{(2)}) > u)}{\mathbb{P}(F^{(2)}(Y^{(2)}) > u)} \equiv \lim_{u \to 1} \chi(u)
\]

- $\chi > 0 \Rightarrow Y^{(1)}$ and $Y^{(2)}$ are AD; the value of $\chi$ 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 \(j\).

![Map of Europe with color gradient indicating extremal correlation values.](image)

**Figure 3:** Estimator of the extremal correlation, \(\hat{\chi}\) between precipitation and wind. \(k = 100\).

\[
\hat{\chi}(a) = \frac{1}{k} \sum_{i=1}^{n} 1_{\{R^{(a,1)}_i > n-k+0.5, R^{(a,2)}_i > n-k+0.5\}}
\]

(1)

where \(R^{(a,\ell)}_i\) denotes the rank of \(Z^{(a,\ell)}_i\) among \(Z^{(a,\ell)}_1, \ldots, Z^{(a,\ell)}_n\), \(\ell = 1, 2\).
Extremal correlation according to distance between two sites

Figure 4: Estimator of the extremal correlation, $\hat{\chi}$ for precipitation data (left) and for wind speed (right)
Dependence-based Regionalisation. Rainfall data.

- Bernard, Naveau, Vrac, Mestre, 2013
  Extremal dependence
  Partitioning 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 $Y_n = (Y_n^{(1)}, \ldots, Y_n^{(d)})$ 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.,

$$\lim_{n \to \infty} \mathbb{P}\left\{ \bigvee_{i=1}^{n} Y_i \leq u_n(x) \right\} = H(x)$$

where $u_n(x)$ 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

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

\[ H = \prod_{j=1}^{d} H^{(j)} \]

- \((\mathbf{Y}_n, n \in \mathbb{N})\) 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

\[ H = \prod_{g=1}^{G} H^{(O_g)} \]
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 $\chi$ 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 \( (Z_n^{(s)}, s \in D \subseteq \mathbb{R}^2, n \in \mathbb{N}) \) be a spatio-temporal random field.
- \( Z_n^{(s)} = (Z_{n}^{(s,1)}, Z_{n}^{(s,2)}) \) 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 \( Z_n = (Z_{n}^{(1)}, \ldots, Z_{n}^{(d)}) \), where \( Z_n \) is a \( 2d \)-random vector with stationary law \( Z = (Z^{(1)}, \ldots, Z^{(d)}) \).
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

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

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

\[ \text{SECO}(Z^{(a)}, Z^{(b)}) = \theta(a) + \theta(b) - \theta(a, b). \]

- 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

\[ \text{SECO}(Z^{(1)}, Z^{(2)}) = 2 - \theta(1,2) = \chi(1,2), \]

if \( Z^{(1)} \) and \( Z^{(2)} \) are univariate random variables.
An empirical estimator of SECO

- The empirical counterpart of the SECO is denoted as \( \hat{\text{SECO}}(Z^{(a)}, Z^{(b)}) \) and is defined as:

\[
\hat{\text{SECO}}(Z^{(a)}, Z^{(b)}) = \hat{\theta}(a) + \hat{\theta}(b) - \hat{\theta}(a, b),
\]

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

\[
\hat{\theta}(j) = \frac{1}{k} \sum_{i=1}^{n} 1_{\{R_{i}^{(j,1)} > n-k+0.5 \text{ or } R_{i}^{(j,2)} > n-k+0.5\}}, \quad j = a, b,
\]

with \( R_{i}^{(j,\ell)} \) the rank of \( Z_{i}^{(j,\ell)} \) among \( Z_{1}^{(j,\ell)}, \ldots, Z_{n}^{(j,\ell)} \), \( j = a, b, \quad \ell = 1, 2. \)
Under **mixing conditions**, we can show that this statistic is consistent i.e.

\[
\sqrt{n} \frac{\text{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)})}{\min\{\hat{\theta}(a), \hat{\theta}(b)\}} \xrightarrow{\mathbb{P}} \text{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)}).
\]

Furthermore, we have

\[
\hat{\Theta}(a, b) := \sqrt{n} \frac{\text{SECO}(\mathbf{Z}^{(a)}, \mathbf{Z}^{(b)})}{\min\{\hat{\theta}(a), \hat{\theta}(b)\}} \in [0, 1].
\]
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
A constrained asymptotic independent block model is defined by

- $Z$ a $2d$-vector with law $F$ having $d$ marginal random vectors (d sites): $Z = (Z^{(1,1)}, Z^{(1,2)}, Z^{(2,1)}, Z^{(2,2)}, \ldots, Z^{(d,1)}, Z^{(d,2)})$
- $F$ is in the max domain of attraction of $H$.
- There exists $O = \{O_g\}_{g=1}^G$ a partition of $\{1, \ldots, d\}$ in $G$ groups with $|O_g| = d_g$ and marginal extreme value distributions $H^{(O_g)}: \mathbb{R}^{d_g} \rightarrow [0, 1]$ such that $H = \prod_{g=1}^G H^{(O_g)}$. 
In a constrained asymptotic independent block model, the following statement

\[ Z^{(O_{g1})} \perp \perp_{\text{ext}} Z^{(O_{g2})} \]

is equivalent to the following statement, implying the SECO

\[ \text{SECO}(a, b) = \text{SECO}(b, a) = 0, \forall a \in O_{g1}, \forall b \in O_{g2}. \]

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
    \[
    \hat{\Theta}(a, b) = \text{SECO}(a, b) / \min\{\hat{\theta}(a), \hat{\theta}(b)\}, \quad a, b \in \{1, \ldots, d\}, \quad (3)
    \]
  - No choice for the number of groups
  - Involving a threshold $\tau$
Algorithm (CAICE) Clustering procedure for AI block models with compound extremes

1: procedure CAICE($S$, $\tau$, $\hat{\Theta}$)
2: Initialize: $S = \{1, \ldots, d\}$, $\hat{\Theta}(a, b)$ for $a, b \in \{1, \ldots, d\}$ and $l = 0$
3: while $S \neq \emptyset$ do
4:     $l = l + 1$
5:     if $|S| > 1$ then
6:         $(a_l, b_l) = \arg \max_{a, b \in S} \hat{\Theta}(a, b)$
7:         if $\hat{\Theta}(a_l, b_l) > \tau$ then
8:             $\hat{O}_l = \{s \in S : \hat{\Theta}(a_l, s) \geq \tau \text{ and } \hat{\Theta}(b_l, s) \geq \tau\}$
9:             if $\hat{\Theta}(a_l, b_l) \leq \tau$ then
10:                $\hat{O}_l = \{a_l\}$
11:         if $|S| = 1$ then
12:             $\hat{O}_l = S$
13:         $S = S \setminus \hat{O}_l$
14:     return $\hat{O} = (\hat{O}_l)_l$
How to set up the threshold $\tau$?

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

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

- Find $\tau$ which minimizes the empirical SECO permit to recover an AI-block model.
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
Calibration of the threshold $\tau$

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

with

$$L(\tau) = \ln \left( 1 + \left( \text{SECO}(X^{(1)}_\tau, \ldots, X^{(G)}_\tau) - \min_{\tau \in \Delta} \text{SECO}(X^{(1)}_\tau, \ldots, X^{(G)}_\tau) \right) \right).$$
Figure 7: Partition of the SECO similarity matrix with threshold $\tau = 0.08$. Squares represent the clusters of variables.
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. $1 - \overline{SECO}$ is used as the dissimilarity matrix.
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
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


- Maume-Deschamps, V., P. Ribereau, and M. Zeidan (2023). Detecting the stationarity of spatial dependence structure using spectral clustering. https://hal.science/hal-03918937