Climate event attribution using multivariate peaks-over-thresholds modelling

Anna Kiriliouk [1] Philippe Naveau [2]

[1]**Université de Namur, Belgium**
[2]Laboratoire des Sciences du Climat et l'Environnement (CNRS), France

26th Annual Meeting of the RSSB, October 18, 2018

Outline

Climate event attribution

- Environmental extreme events: increase in frequency and magnitude in view of climate change.
- Climate scientists want to measure the impact of humans on changes in the earth's climate system
  $\rightarrow$ **event attribution.**
- Use climate models to compare probabilities of an extreme event with and without **anthropogenic forcing**.

By how much have human activities increased the risk of occurrence of an extreme event?

- Compare probability of an extreme event in the factual world (p_1) and in a counterfactual world (p_0).
  - Counterfactual world: without anthropogenic forcing, i.e., without human influence.
  - Extreme event: a random variable of interest (rainfall, temperature, ...) exceeds some high threshold.

By how much have human activities increased the risk of occurrence of an extreme event?

- Compare probability of an extreme event in the factual world (p_1) and in a counterfactual world (p_0).
  - Counterfactual world: without anthropogenic forcing, i.e., without human influence.
  - Extreme event: a random variable of interest (rainfall, temperature, ...) exceeds some high threshold.
- The **Fraction of Attributable Risk (FAR)** (Stott et al., 2004) is

$$\text{FAR} = \frac{p_1 - p_0}{p_1} = 1 - \frac{p_0}{p_1}.$$

By how much have human activities increased the risk of occurrence of an extreme event?

- Compare probability of an extreme event in the factual world (p_1) and in a counterfactual world (p_0).
  - Counterfactual world: without anthropogenic forcing, i.e., without human influence.
  - Extreme event: a random variable of interest (rainfall, temperature, ...) exceeds some high threshold.
- The **Fraction of Attributable Risk (FAR)** (Stott et al., 2004) is

$$\mathrm{FAR} = \frac{p_1 - p_0}{p_1} = 1 - \frac{p_0}{p_1}.$$

- Many attribution studies are based on the normal distribution (not heavy-tailed), or use nonparametric estimation (no extrapolation possible), or are computationally heavy.
- We use **multivariate peaks-over-thresholds modelling** (K., Rootzén, Segers, and Wadsworth, 2018) to characterize p_0 and p_1.

Probabilities of causation of extreme events

- Let C denote a cause (*human activities*) and E an event (*a climatic index exceeding a high threshold*).
- Three distinct forms of causation can be of interest (Pearl, 2000)

Probabilities of causation of extreme events

- Let C denote a cause (*human activities*) and E an event (*a climatic index exceeding a high threshold*).
- Three distinct forms of causation can be of interest (Pearl, 2000)
  - PN Probability of necessary causation: C is required for E to occur but other factors may be required as well.
  - PS Probability of sufficient causation: C always triggers E, but E might occur without C.
  - PNS Probability of necessary and sufficient causation.

Probabilities of causation of extreme events

- Let C denote a cause (*human activities*) and E an event (*a climatic index exceeding a high threshold*).
- Three distinct forms of causation can be of interest (Pearl, 2000)
  - **PN** Probability of necessary causation: C is required for E to occur but other factors may be required as well.
  - **PS** Probability of sufficient causation: C always triggers E, but E might occur without C.
  - **PNS** Probability of necessary and sufficient causation.
- Under certain conditions (Hannart et al., 2016), these probabilities can be calculated by

$$\mathrm{PN} = 1 - \frac{p_0}{p_1}, \quad \mathrm{PS} = 1 - \frac{1-p_1}{1-p_0}, \quad \mathrm{PNS} = p_1 - p_0.$$

Probabilities of causation of extreme events

- Let C denote a cause (*human activities*) and E an event (*a climatic index exceeding a high threshold*).
- Three distinct forms of causation can be of interest (Pearl, 2000)
  - PN Probability of necessary causation: C is required for E to occur but other factors may be required as well.
  - PS Probability of sufficient causation: C always triggers E, but E might occur without C.
  - PNS Probability of necessary and sufficient causation.
- Under certain conditions (Hannart et al., 2016), these probabilities can be calculated by

$$\text{PN} = 1 - \frac{p_0}{p_1}, \quad \text{PS} = 1 - \frac{1-p_1}{1-p_0}, \quad \text{PNS} = p_1 - p_0.$$

- The FAR coincides with the probability that human activities are necessary for the event to occur.
- When this threshold is high, the event is rare in the factual world (p_1 small) and nearly impossible in the counterfactual worls ($p_0 \approx 0$).

Outline

Univariate peaks over thresholds modelling

- The distribution function H of a **generalized Pareto distribution** is

$$H(x; \sigma, \gamma) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)_+^{-1/\gamma},$$

where $\sigma > 0$ and $\gamma \in \mathbb{R}$ are the scale and shape parameters.

Univariate peaks over thresholds modelling

- The distribution function H of a **generalized Pareto distribution** is

$$H(x;\sigma,\gamma) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)_+^{-1/\gamma},$$

  where $\sigma > 0$ and $\gamma \in \mathbb{R}$ are the scale and shape parameters.
- If Y is a random variable and u a sufficiently high threshold,

$$Y - u \mid Y > u \overset{d}{\approx} X,$$

  where $X \sim \text{GPD}(\sigma,\gamma)$ (Pickands-Balkema-de Haan theorem).
- The parameter γ characterizes the heaviness of the tail and the support of the distribution.

Univariate peaks over thresholds modelling

- The distribution function H of a **generalized Pareto distribution** is

$$H(x;\sigma,\gamma) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)_+^{-1/\gamma},$$

where $\sigma > 0$ and $\gamma \in \mathbb{R}$ are the scale and shape parameters.

- If Y is a random variable and u a sufficiently high threshold,

$$Y - u \mid Y > u \stackrel{d}{\approx} X,$$

where $X \sim \text{GPD}(\sigma, \gamma)$ (Pickands-Balkema-de Haan theorem).

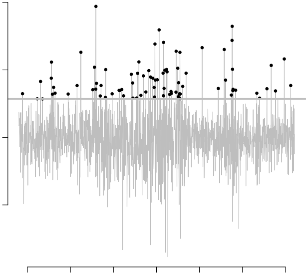
- The parameter γ characterizes the heaviness of the tail and the support of the distribution.
- The GPD can be used to model the tail of Y, since for $y > u$,

$$\mathbb{P}[Y > y] \approx \mathbb{P}[Y > u]\, \overline{H}(y - u; \sigma, \gamma).$$

Univariate vs multivariate peaks-over-thresholds

Univariate modelling strategy:

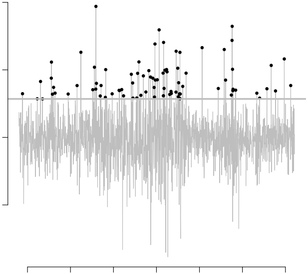
- Fix some high threshold u.
- Fit a GPD to the conditional threshold excesses $Y - u \mid Y > u$.



Univariate vs multivariate peaks-over-thresholds

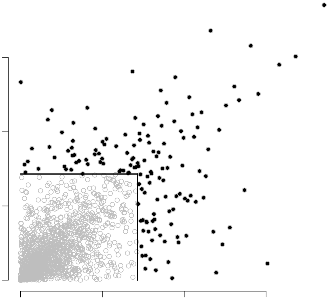
Univariate modelling strategy:

- Fix some high threshold u.
- Fit a GPD to the conditional threshold excesses $Y - u \mid Y > u$.



Multivariate modelling strategy:

- Fix some high threshold $\boldsymbol{u}$.
- Fit a multivariate GPD to the conditional threshold excesses $\boldsymbol{Y} - \boldsymbol{u} \mid \boldsymbol{Y} \not\leq \boldsymbol{u}$.



Characterizing a multivariate GPD

- Let $\boldsymbol{Y}$ be a d-dimensional random vector and $\boldsymbol{u} \in \mathbb{R}^d$ a threshold.
- If $\boldsymbol{X} \approx \boldsymbol{Y} - \boldsymbol{u} \mid \boldsymbol{Y} \not\leq \boldsymbol{u}$, then $\boldsymbol{X}$ follows approximately a multivariate GPD (Rootzén and Tajvidi, 2006).
- The distribution function H of a multivariate GPD vector $\boldsymbol{X}$ belongs to an infinite-dimensional parametric family: a model construction tool has been proposed in Rootzén et al. (2018).

Characterizing a multivariate GPD

- Let $\boldsymbol{Y}$ be a d-dimensional random vector and $\boldsymbol{u} \in \mathbb{R}^d$ a threshold.
- If $\boldsymbol{X} \approx \boldsymbol{Y} - \boldsymbol{u} \mid \boldsymbol{Y} \not\leq \boldsymbol{u}$, then $\boldsymbol{X}$ follows approximately a multivariate GPD (Rootzén and Tajvidi, 2006).
- The distribution function H of a multivariate GPD vector $\boldsymbol{X}$ belongs to an infinite-dimensional parametric family: a model construction tool has been proposed in Rootzén et al. (2018).
- The margins of a multivariate GPD are generally **not** GPDs, but the conditional margins are GPDs, i.e.,

$$\mathbb{P}[X_j \leq x \mid X_j > 0] = H(x; \sigma_j, \gamma_j), \qquad j \in \{1, \ldots, d\}.$$

Characterizing a multivariate GPD

- Let $\boldsymbol{Y}$ be a d-dimensional random vector and $\boldsymbol{u} \in \mathbb{R}^d$ a threshold.
- If $\boldsymbol{X} \approx \boldsymbol{Y} - \boldsymbol{u} \mid \boldsymbol{Y} \nleq \boldsymbol{u}$, then $\boldsymbol{X}$ follows approximately a multivariate GPD (Rootzén and Tajvidi, 2006).
- The distribution function H of a multivariate GPD vector $\boldsymbol{X}$ belongs to an infinite-dimensional parametric family: a model construction tool has been proposed in Rootzén et al. (2018).
- The margins of a multivariate GPD are generally **not** GPDs, but the conditional margins are GPDs, i.e.,

$$\mathbb{P}[X_j \leq x \mid X_j > 0] = H(x; \sigma_j, \gamma_j), \qquad j \in \{1, \ldots, d\}.$$

- Suppose that $\gamma := \gamma_1 = \ldots = \gamma_d$, i.e., the shape parameters of the conditional margins are equal. For weights $\boldsymbol{w} = (w_1, \ldots, w_d) > \boldsymbol{0}$,

$$\boldsymbol{w}^T\boldsymbol{X} \mid \boldsymbol{w}^T\boldsymbol{X} > 0 \sim \mathrm{GPD}\left(\boldsymbol{w}^T\boldsymbol{\sigma}, \gamma\right).$$

Outline

Modelling causation probabilities: univariate case

- Let $Y^{(0)}$, $Y^{(1)}$ denote the climate model output in the counterfactual and factual world respectively.
- Assume $Y^{(i)} - u^{(i)} \mid Y^{(i)} > u^{(i)} \approx X^{(i)} \sim \text{GPD}(\sigma^{(i)}, \gamma^{(i)})$ for $i = 0, 1$.

Modelling causation probabilities: univariate case

- Let $Y^{(0)}$, $Y^{(1)}$ denote the climate model output in the counterfactual and factual world respectively.
- Assume $Y^{(i)} - u^{(i)} \mid Y^{(i)} > u^{(i)} \approx X^{(i)} \sim \mathrm{GPD}(\sigma^{(i)}, \gamma^{(i)})$ for $i = 0, 1$.
- Set

$$\mathrm{FAR}(v) = \mathrm{PN}(v) = 1 - \frac{p_0(v)}{p_1(v)},$$

  where $p_i(v) = \mathbb{P}[Y^{(i)} > v]$.
- For $v > u^{(i)}$,

$$p_i(v) \approx \mathbb{P}\left[Y^{(i)} > u^{(i)}\right] \overline{H}\left(v - u^{(i)}; \sigma^{(i)}, \gamma^{(i)}\right).$$

Modelling causation probabilities: univariate case

- Let $Y^{(0)}$, $Y^{(1)}$ denote the climate model output in the counterfactual and factual world respectively.
- Assume $Y^{(i)} - u^{(i)} \mid Y^{(i)} > u^{(i)} \approx X^{(i)} \sim \mathrm{GPD}(\sigma^{(i)}, \gamma^{(i)})$ for $i = 0, 1$.
- Set

$$\mathrm{FAR}(v) = \mathrm{PN}(v) = 1 - \frac{p_0(v)}{p_1(v)},$$

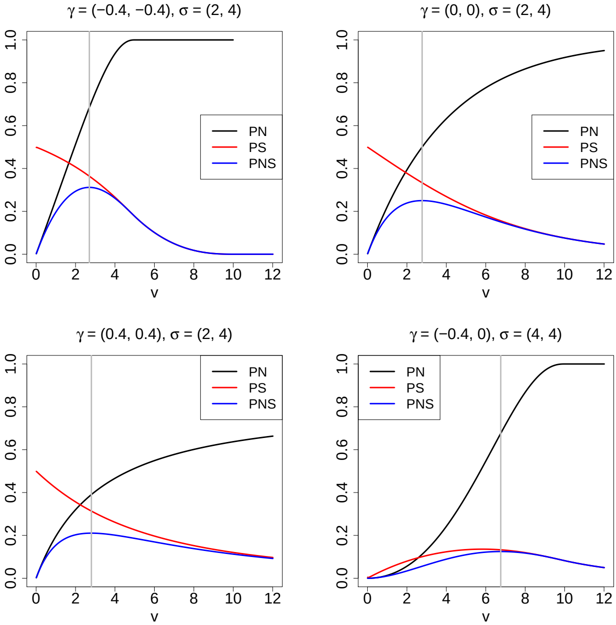
where $p_i(v) = \mathbb{P}[Y^{(i)} > v]$.

- For $v > u^{(i)}$,

$$p_i(v) \approx \mathbb{P}\left[Y^{(i)} > u^{(i)}\right] \overline{H}\left(v - u^{(i)}; \sigma^{(i)}, \gamma^{(i)}\right).$$

- The probability $\mathbb{P}[Y^{(i)} > u^{(i)}]$ can be estimated by the empirical excess probability of $u^{(i)}$.
- Estimating p_0 and p_1 for high values of v then amounts to estimating the parameters of a univariate GPD.

The PN, PS and PNS for different GPD parameters



Modelling causation probabilities: multivariate case

- Let $\boldsymbol{Y}^{(0)}$, $\boldsymbol{Y}^{(1)} \in \mathbb{R}^d$ be random vectors of climate model output in the counterfactual and factual world respectively.
- Assume $\boldsymbol{Y}^{(i)} - \boldsymbol{u}^{(i)} \mid \boldsymbol{Y}^{(i)} \nleq \boldsymbol{u}^{(i)} \approx \boldsymbol{X}^{(i)}$ for some high threshold $\boldsymbol{u}^{(i)}$, where $\boldsymbol{X}^{(i)}$ is a multivariate GPD vector.

Modelling causation probabilities: multivariate case

- Let $\boldsymbol{Y}^{(0)}$, $\boldsymbol{Y}^{(1)} \in \mathbb{R}^d$ be random vectors of climate model output in the counterfactual and factual world respectively.
- Assume $\boldsymbol{Y}^{(i)} - \boldsymbol{u}^{(i)} \mid \boldsymbol{Y}^{(i)} \not\leq \boldsymbol{u}^{(i)} \approx \boldsymbol{X}^{(i)}$ for some high threshold $\boldsymbol{u}^{(i)}$, where $\boldsymbol{X}^{(i)}$ is a multivariate GPD vector.
- Set

$$\mathrm{PN}(v; \boldsymbol{w}) = \mathrm{FAR}(v; \boldsymbol{w}) = 1 - \frac{p_0(v; \boldsymbol{w})}{p_1(v; \boldsymbol{w})},$$

for weights $\boldsymbol{w} = (w_1, \ldots, w_d)^T > \mathbf{0}$ with $\sum_{j=1}^d w_j = 1$, and

$$p_i(v) = \mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{Y}^{(i)} > v\right].$$

Modelling causation probabilities: multivariate case

- Let $\boldsymbol{Y}^{(0)}$, $\boldsymbol{Y}^{(1)} \in \mathbb{R}^d$ be random vectors of climate model output in the counterfactual and factual world respectively.
- Assume $\boldsymbol{Y}^{(i)} - \boldsymbol{u}^{(i)} \mid \boldsymbol{Y}^{(i)} \not\leq \boldsymbol{u}^{(i)} \approx \boldsymbol{X}^{(i)}$ for some high threshold $\boldsymbol{u}^{(i)}$, where $\boldsymbol{X}^{(i)}$ is a multivariate GPD vector.
- Set

$$\mathrm{PN}(v; \boldsymbol{w}) = \mathrm{FAR}(v; \boldsymbol{w}) = 1 - \frac{p_0(v; \boldsymbol{w})}{p_1(v; \boldsymbol{w})},$$

for weights $\boldsymbol{w} = (w_1, \ldots, w_d)^T > \boldsymbol{0}$ with $\sum_{j=1}^d w_j = 1$, and

$$p_i(v) = \mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{Y}^{(i)} > v\right].$$

- Recall the sum-stability property,

$$\mathbb{P}[\boldsymbol{w}^T \boldsymbol{X}^{(i)} > v] = \mathbb{P}[\boldsymbol{w}^T \boldsymbol{X}^{(i)} > 0]\,\overline{H}(v; \boldsymbol{w}^T \boldsymbol{\sigma}^{(i)}, \gamma^{(i)}).$$

Modelling causation probabilities: multivariate case

- The sum-stability property is only valid when $\gamma^{(i)} := \gamma_1^{(i)} = \ldots = \gamma_d^{(i)}$ for $i \in \{0, 1\}$.
- Combining $\boldsymbol{Y}^{(i)} - \boldsymbol{u}^{(i)} \mid \boldsymbol{Y}^{(i)} \not\leq \boldsymbol{u}^{(i)} \approx \boldsymbol{X}^{(i)}$ and sum-stability, we find

$$p_i(v) \approx \mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{Y}^{(i)} > \boldsymbol{w}^T \boldsymbol{u}^{(i)}\right] \overline{H}\left(v - \boldsymbol{w}^T \boldsymbol{u}^{(i)}; \boldsymbol{w}^T \boldsymbol{\sigma}^{(i)}, \gamma^{(i)}\right)$$

for $v > \boldsymbol{w}^T \boldsymbol{u}^{(i)}$.

Modelling causation probabilities: multivariate case

- The sum-stability property is only valid when $\gamma^{(i)} := \gamma_1^{(i)} = \ldots = \gamma_d^{(i)}$ for $i \in \{0, 1\}$.
- Combining $\boldsymbol{Y}^{(i)} - \boldsymbol{u}^{(i)} \mid \boldsymbol{Y}^{(i)} \not\leq \boldsymbol{u}^{(i)} \approx \boldsymbol{X}^{(i)}$ and sum-stability, we find

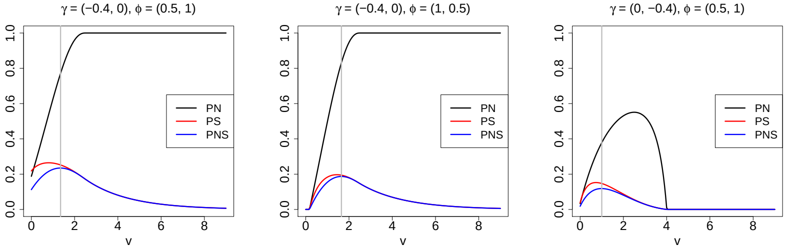
$$p_i(v) \approx \mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{Y}^{(i)} > \boldsymbol{w}^T \boldsymbol{u}^{(i)}\right] \overline{H}\left(v - \boldsymbol{w}^T \boldsymbol{u}^{(i)}; \boldsymbol{w}^T \boldsymbol{\sigma}^{(i)}, \gamma^{(i)}\right)$$

  for $v > \boldsymbol{w}^T \boldsymbol{u}^{(i)}$.
- The first term on the right-hand side can be estimated non-parametrically.
- The parameters σ_j, γ can be estimated by fitting GPDs to the conditional marginal threshold exceedances.

The PN, PS and PNS for different scenarios

Examples of causation probabilities in $d = 2$ with $\boldsymbol{w} = (1/2, 1/2)^T$ and $\gamma = \gamma^{(0)} = \gamma^{(1)}$, where, from left to right,

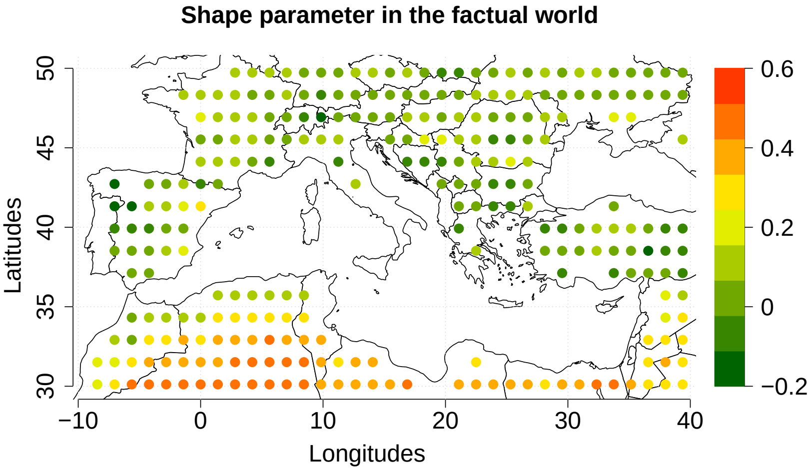
1. The factual world exhibits increased dependence and increased marginal tail heaviness.
2. The factual world shows decreased dependence and increased marginal tail heaviness.
3. The factual world shows increased dependence and decreased marginal tail heaviness.



Outline

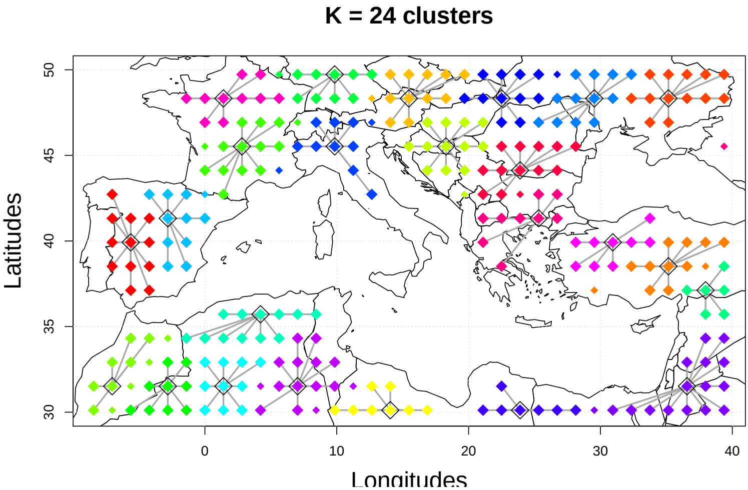
European precipitation: shape parameters of GPD fit

Weekly maximum precipitation amounts (mm/day) between 1850 and 2005 on a grid of $d = 268$ points in central Europe.



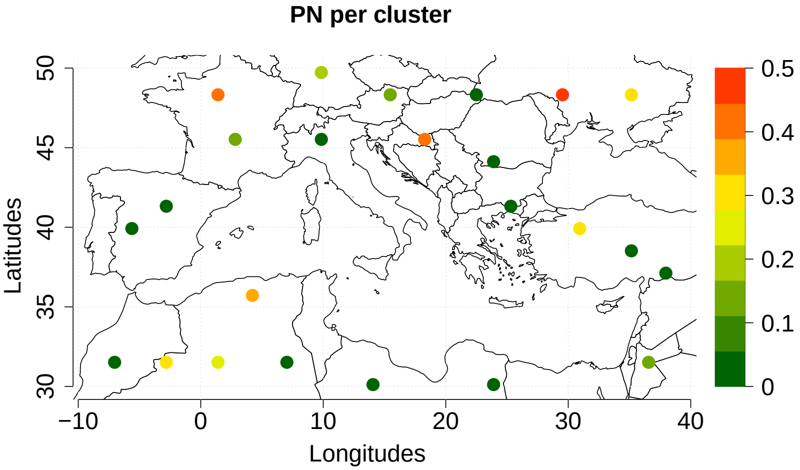
European precipitation: clusters

Partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990) based on madogram distance.



European precipitation: probabilities of necessary causation

Fraction of Attributable Risk (or PN) per cluster, evaluated at a threshold v corresponding to a **hundred-year return level** of $\boldsymbol{w}^T \boldsymbol{Y}^{(1)}$.



Discussion

- For $v > \boldsymbol{w}^T \boldsymbol{u}$, we can also use the approximation

$$\mathbb{P}[\boldsymbol{w}^T \boldsymbol{Y} > v] \approx \mathbb{P}[\boldsymbol{Y} \not\leq \boldsymbol{u}]\, \mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{X} > 0\right] \overline{H}\left(v - \boldsymbol{w}^T \boldsymbol{u}; \boldsymbol{w}^T \boldsymbol{\sigma}, \gamma\right).$$

- For certain parametric models on the multivariate GPD vector $\boldsymbol{X}$ and for certain γ, the probability $\mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{X} > 0\right]$ can be calculated analytically.

Discussion

- For $v > \boldsymbol{w}^T \boldsymbol{u}$, we can also use the approximation

$$\mathbb{P}[\boldsymbol{w}^T \boldsymbol{Y} > v] \approx \mathbb{P}[\boldsymbol{Y} \not\leq \boldsymbol{u}] \, \mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{X} > 0\right] \overline{H}\left(v - \boldsymbol{w}^T \boldsymbol{u}; \boldsymbol{w}^T \boldsymbol{\sigma}, \gamma\right).$$

- For certain parametric models on the multivariate GPD vector $\boldsymbol{X}$ and for certain γ, the probability $\mathbb{P}\left[\boldsymbol{w}^T \boldsymbol{X} > 0\right]$ can be calculated analytically.
- To enhance the FAR, we look for weights $w_1, \ldots, w_d$ that are optimal, i.e.,

$$\boldsymbol{w}_{\text{opt}} = \underset{\boldsymbol{w}: \sum_j w_j = 1}{\arg\max} \; \text{FAR}(v; \boldsymbol{w}).$$

$\rightarrow$ analytic expression for $\boldsymbol{w}_{\text{opt}}$ only available in $d = 2$ for special cases.

References

Hannart, A., J. Pearl, F. Otto, P. Naveau, and M. Ghil (2016). Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society 97*(1), 99–110.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data: an introduction to cluster analysis*. Wiley Online Library.

Kiriliouk, A. and P. Naveau (2018). Climate event attribution using peaks-over-thresholds modelling. In preparation.

Kiriliouk, A., H. Rootzén, J. Segers, and J. L. Wadsworth (2018). Peaks over thresholds modelling with multivariate generalized Pareto distributions. To be published in Technometrics.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge University Press.

Rootzén, H., J. Segers, and J. L. Wadsworth (2018). Multivariate peaks over thresholds models. *Extremes 21*(1), 115–145.

Rootzén, H. and N. Tajvidi (2006). Multivariate generalized Pareto distributions. *Bernoulli 12*(5), 917–930.

Stott, P. A., D. A. Stone, and M. R. Allen (2004). Human contribution to the european heatwave of 2003. *Nature 432*(7017), 610.