Causal Discovery in Multivariate Extremes with a Hydrological Analysis of Swiss River Discharges

V. Chavez-Demoulin[1] L. Mhalla[2] P. Naveau[3]

[1]HEC, University of Lausanne, Switzerland

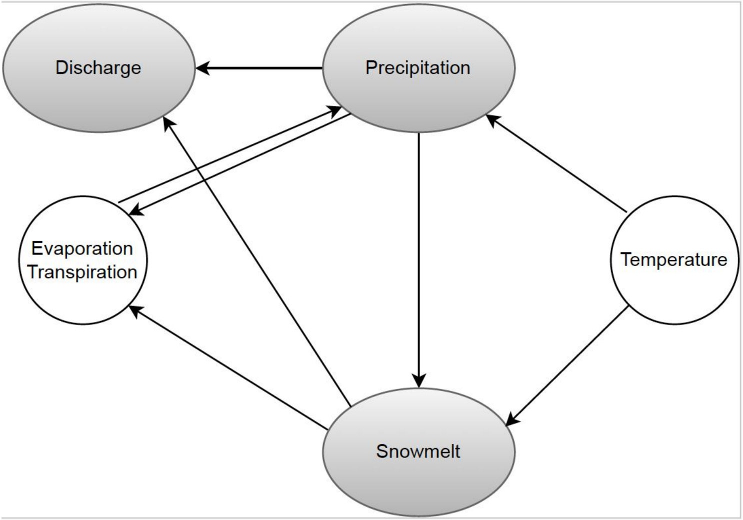
[2]Institute of Mathematics, EPFL, Switzerland

[3]CNRS LSCE, France

1 Motivation

Goal: Retrieve causal information from observational data, but at **extreme levels**

- The generating mechanism of the system of interest might behave differently in the bulk than in the tails
- In climate models, often there might be cycles



2 Related Work

- Gnecco et al. (2021): heavy-tailed linear LSCM

$$Y_j := \sum_{k \in \mathrm{pa}(j)} \beta_{jk} Y_k + \varepsilon_j, \quad \text{node } j$$

- the generating mechanism propagates into the tails
- extreme events in a node are determined by the noise terms of its ancestors

$\Rightarrow$ some climate mechanisms cannot be represented under this structure, e.g., precipitation and soil moisture

- Gissibl and Klüppelberg (2018): recursive max-linear models (RMLM)

$$Y_j := \max_{k \in \mathrm{pa}(j)} \max\left(\beta_{jk} Y_k, \varepsilon_j\right), \quad \text{node } j$$

- a node is the maximum shock among a set of independent heavy-tailed factors
- the resulting joint distributions are discrete $\Rightarrow$ tricky to work with

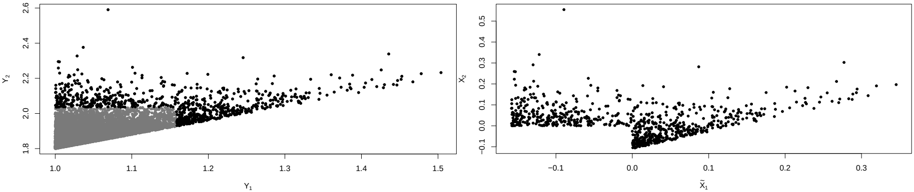
3 Setting

Assumptions:

- $\mathbf{Y} = (Y_1, \ldots, Y_d)^\top \sim F$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ i.i.d. copies
- F in the max-domain of attraction of a max-stable distribution G
- $$\mathbf{Y} - \mathbf{u} \,\Big|\, \mathbf{Y} \not\leq \mathbf{u} \xrightarrow{d} \widetilde{\mathbf{X}} \quad \text{as } \mathbf{u} \to \infty$$

$\Rightarrow \widetilde{\mathbf{X}}$ follows a multivariate GP distribution associated with G

Example: Bivariate heavy-tailed LSCM



- We remove the marginal information and work with $\mathbf{X}$ s.t.

$$\widetilde{\mathbf{X}} = \sigma \frac{e^{\xi \mathbf{X}} - 1}{\xi},$$

with σ and ξ the marginal GP parameters

$\Rightarrow \mathbf{X}$ is a standard Pareto random vector and can be expressed through its spectral representation (Ferreira and Haan 2014)

$$\mathbf{X} = E + \mathbf{U} - \max(\mathbf{U}),$$

where $\mathbf{U} = (U_1, \ldots, U_d)^\top$ independent of $E \sim Exp(1)$

Examples:

- U_i independent Gumbel r.v.s with equal scale yield the MGPD associated with the logistic max-stable distribution
- U_i independent log-gamma r.v.s yield the MGPD associated with the Dirichlet max-stable distribution

5 Setting

Equivalently

- the standard bivariate Pareto vector has the following construction

$$\begin{cases} X_1 = E + V - \max(V, 0) \\ X_2 = E - \max(V, 0) \end{cases}$$

  with $V (= U_1 - U_2) \perp\!\!\!\perp E$

- the standard trivariate Pareto vector has the following construction

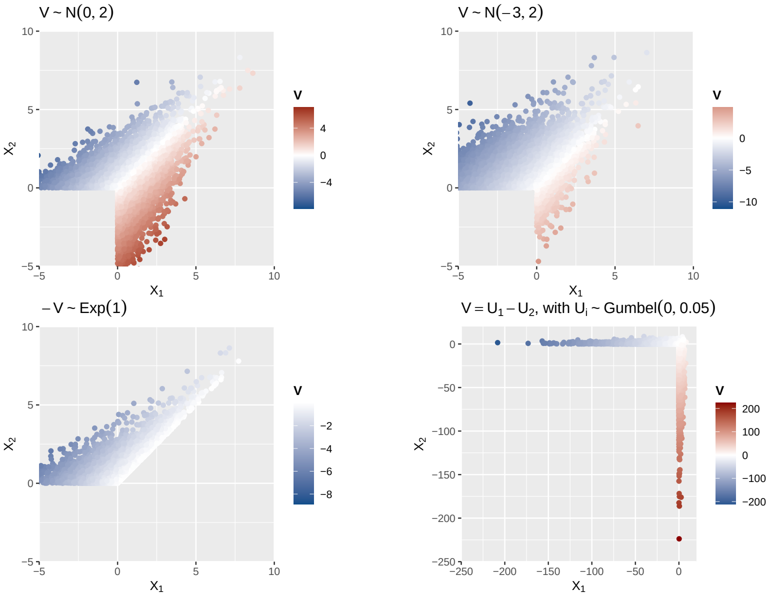
$$\begin{cases} X_1 = E + V_1 - \max(V_1, V_2, 0) \\ X_2 = E + V_2 - \max(V_1, V_2, 0) \\ X_3 = E - \max(V_1, V_2, 0) \end{cases}$$

  with $V_1 (= U_1 - U_3), V_2 (= U_2 - U_3) \perp\!\!\!\perp E$

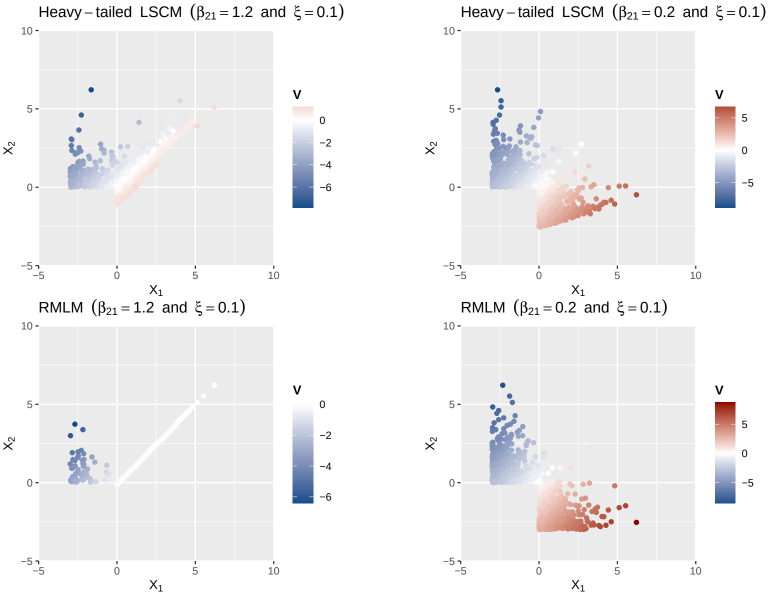
$\Rightarrow$ the quantity $\max(\mathbf{V}, 0)$ plays a key role in the system

- it drives the dependence between the nodes
- it will be central to our definition of **extremal causality**

6 Bivariate Examples: Dependence



7 Bivariate Examples: Causality



8 Extremal Causality: Useful Tools

- The Wasserstein distance (of order 1) between univariate random variables $X_1 \sim F_1$ and $X_2 \sim F_2$ is defined as

$$W(X_1, X_2) = \int_{\mathbb{R}} |F_1(t) - F_2(t)| dt$$

- For $\mathbf{X} = (X_1, \ldots, X_d)^\top$ a standard Pareto r.v., we have

$$W(X_i, E) \geq W(X_j, E) \Leftrightarrow \mathbb{E}(X_i) \leq \mathbb{E}(X_j)$$

where $E \sim Exp(1)$ that dominates X.

9 Extremal Causality: Definition

- $\mathbf{X}$ a standard Pareto r.v.
- Define

$$d_{i \to j} = \{W(X_i, E) - W(X_j, E)\} / \max_k \{W(X_k, E)\}$$

the causal score between X_i and X_j

Definition

X_i is the extremal cause of X_j, whenever $d_{i \to j}$ is finite and > 0

E.g., in the bivariate case, $d_{i \to j} = -\mathbb{E}(V)$

- Causal links as encoded in a heavy-tailed SEM or max-linear model remain valid at extremal levels, according to our definition
- Higher (absolute) values of $d_{i \to j}$ reflect stronger causal links. But not too strong ...

10 Extremal Causality: Remarks

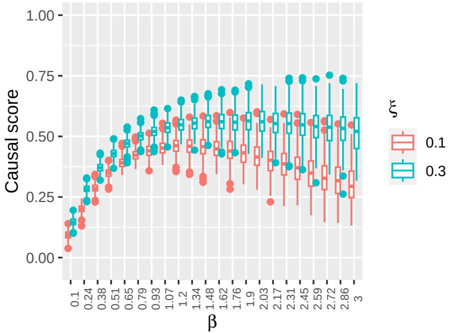
- Asymmetric and strong extremal dependence reflects a strong extremal causal link, even when there is no apparent structure of causality between the variables
- Asymmetric and weak extremal dependence reflects a weak extremal causal link
- Strong/weak extremal dependence coupled with symmetry reflects absence of extremal causal link

$\Rightarrow$ Our definition of **extremal causality** stems from the strength and asymmetry of the extremal dependence structure

- Heavy-tailed LSCM

$$\begin{cases} Y_1 = \epsilon_1 \\ Y_2 = \beta Y_1 + \epsilon_2 \end{cases}$$

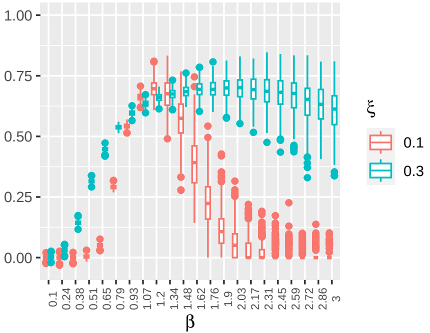
with $\epsilon_i \sim$ Pareto with ξ=0.1



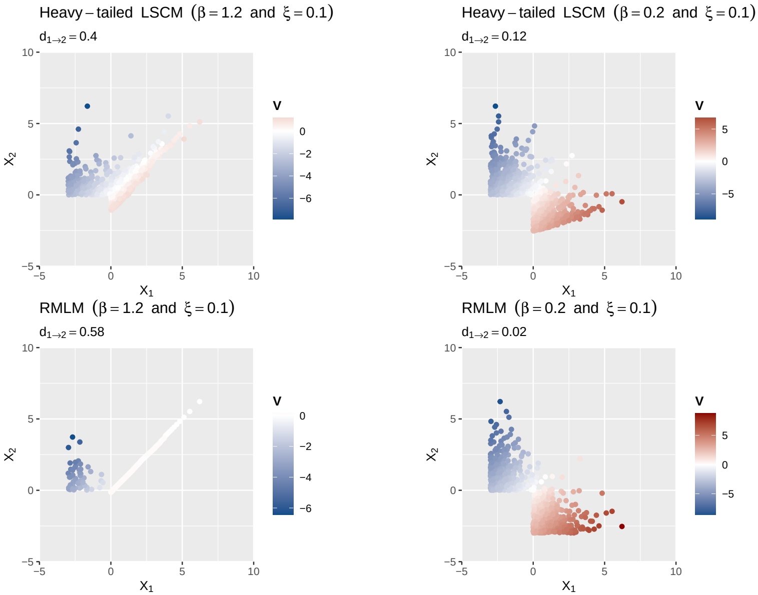
- RMLM

$$\begin{cases} Y_1 = \epsilon_1 \\ Y_2 = \max(\beta Y_1, \epsilon_2) \end{cases}$$

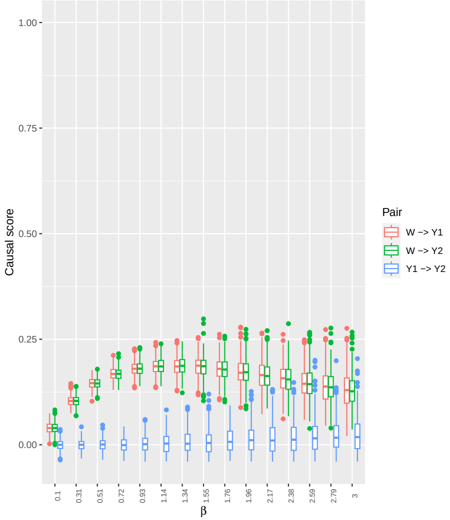
with $\epsilon_i \sim$ Pareto with ξ=0.1

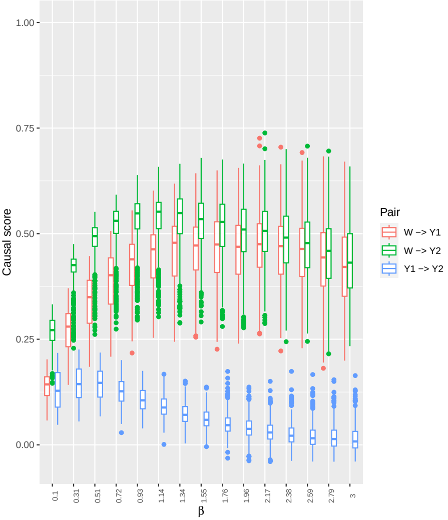


12 Simulations (SEM illustration)

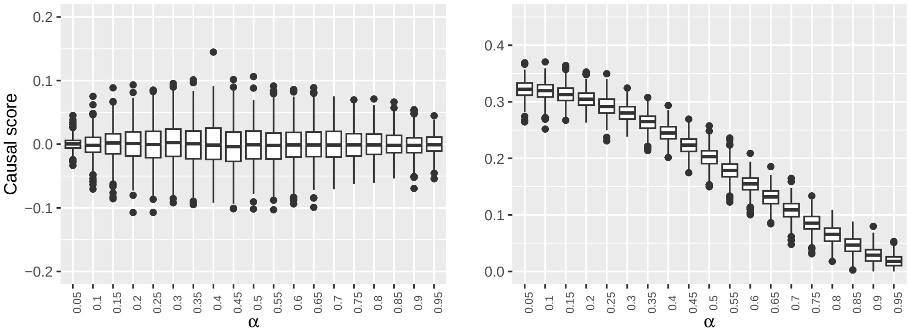


13 Simulations (SEM with confounder)





14 Simulations: (asym) logistic model



15 Floods in Switzerland: Hydrologically simulated dataset

- Data simulated using the hydrological modelling system PREVAH (PREecipitation-Runoff-EVApotranspiration Hydrotope model) (Brunner et al. 2019; Viviroli et al. 2009)
- Dataset consists of 307 catchments in Switzerland for which
  - discharge
  - precipitation
  - snowmelt

were simulated at a daily-resolution from 1981 to 2016

$\Rightarrow$ system of the hydrological variables is spatially dynamic

16 Floods in Switzerland: Extreme hydrological drivers

Goal: assess the extremal causal mechanisms over all catchments during Spring (March-April-May)

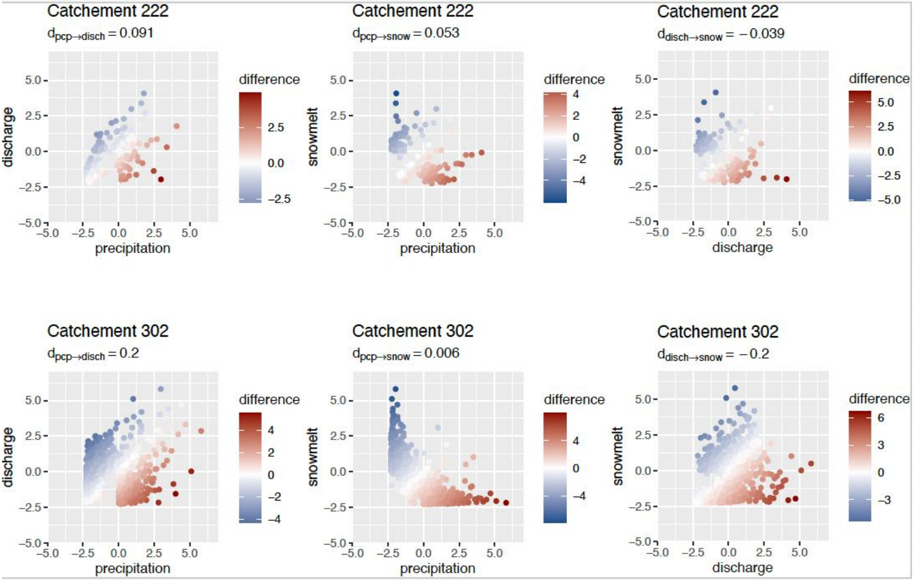
Data pre-processing: take into account the time-lag between the hydrological variables

$\Rightarrow$ temporally match the variables without imposing a time direction that might bias their causal dynamics ($n = 3309$ obs.)

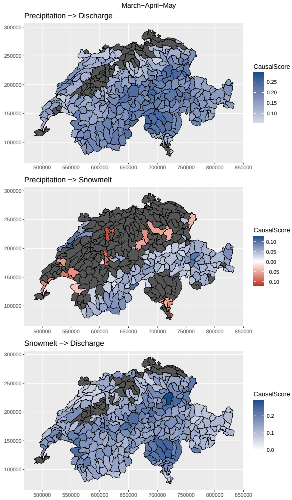
For each of the 307 catchments

- we compute the pairwise causal scores $d_{i \to j}$, for $i, j = 1, \dots, 3 (i \neq j)$
- we bootstrap the data to assess uncertainty of $d_{i \to j}$ ($B = 300$)
  $\Rightarrow$ retain causal scores with $0 \notin CI$

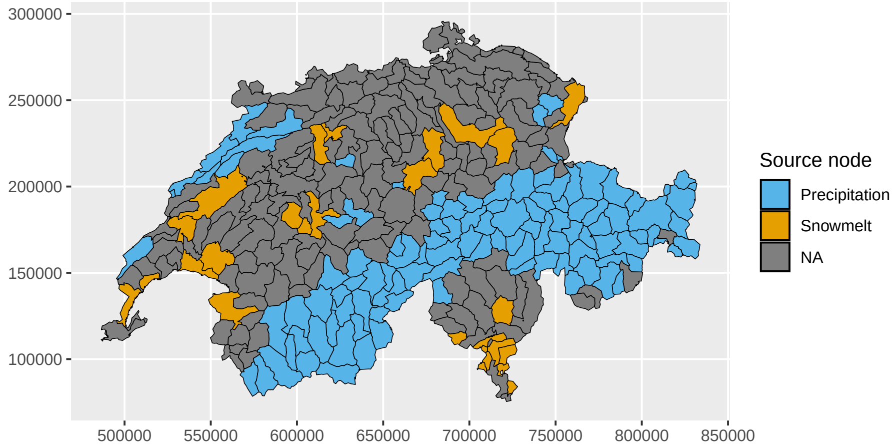
17 Floods in Switzerland: Extreme hydrological drivers



18 Floods in Switzerland: Extreme hydrological drivers



19 Floods in Switzerland: Extreme hydrological drivers



20 Conclusions

- We define **extremal causality** relying on a notion of asymmetry in the limiting MGPD
- In the multivariate setting: $d_{i \to j} > 0$, $\forall j$ is equivalent to i being a source node

$\Rightarrow$ the topological (causal) order is retrieved by ordering $W(X_i, E)$

20 Conclusions

- We define **extremal causality** relying on a notion of asymmetry in the limiting MGPD
- In the multivariate setting: $d_{i \to j} > 0$, $\forall j$ is equivalent to i being a source node

$\Rightarrow$ the topological (causal) order is retrieved by ordering $W(X_i, E)$

Outlook

- One can look at causal links in the U_is of the spectral representation of the MGPD $\Rightarrow$ the system can then be seen as a directed acyclic mixed graph (Henckel et al. 2023)
- If we manage to have the spectral representation of the MGPD associated to the Hüsler–Reiss model, can we link the notion of conditional independence (Engelke and Hitz 2020) to our definition of extremal causality?

21 References

Thank you!

Beljaars, A., Viterbo, P., Miller, M., and Betts, A. (1996), "The anomalous rainfall over the United States during July 1993: Sensitivity to land surface parameterization and soil moisture anomalies," *Monthly Weather Review*, 124, 362–383. https://doi.org/10.1175/1520-0493(1996)124%3C0362:TAROTU%3E2.0.CO;2.

Brunner, M. I., Hingray, B., Zappa, M., and Favre, A.-C. (2019), "Future trends in the interdependence between flood peaks and volumes: Hydro-climatological drivers and uncertainty," *Water Resources Research*, 55, 4745–4759. https://doi.org/https://doi.org/10.1029/2019WR024701.

Engelke, S., and Hitz, A. S. (2020), "Graphical Models for Extremes," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 871–932. https://doi.org/10.1111/rssb.12355.

Ferreira, A., and Haan, L. de (2014), "The generalized Pareto process; with a view towards application and simulation," *Bernoulli*, 20, 1717–1737. https://doi.org/10.3150/13-BEJ538.

Gissibl, N., and Klüppelberg, C. (2018), "Max-linear models on directed acyclic graphs," *Bernoulli*, 24, 2693–2720. https://doi.org/10.3150/17-BEJ941.

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. (2021), "Causal discovery in heavy-tailed models," *The Annals of Statistics*, 49, 1755–1778. https://doi.org/10.1214/20-AOS2021.

Henckel, L., Buttenschoen, M., and Maathuis, M. H. (2023), "Graphical tools for selecting conditional instrumental sets," *Biometrika*. https://doi.org/10.1093/biomet/asad066.

Koukoula, M., Schwartz, C. S., Nikolopoulos, E. I., and Anagnostou, E. N. (2021), "Understanding the impact of soil moisture on precipitation under different climate and meteorological conditions: A numerical sensitivity study over the CONUS," *Journal of Geophysical Research: Atmospheres*, 126, e2021JD035096. https://doi.org/https://doi.org/10.1029/2021JD035096.

Viviroli, D., Zappa, M., Gurtz, J., and Weingartner, R. (2009), "An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools," *Environmental Modelling & Software*, 24, 1209–1222. https://doi.org/https://doi.org/10.1016/j.envsoft.2009.04.001.