Extremes (2023) 26:573–594
https://doi.org/10.1007/s10687-022-00456-4

Causal modelling of heavy-tailed variables and confounders with application to river flow

Olivier C. Pasche[1,3] · Valérie Chavez-Demoulin[2] · Anthony C. Davison[3]

Received: 9 November 2021 / Revised: 21 September 2022 / Accepted: 5 November 2022 /
Published online: 17 December 2022

Abstract
Confounding variables are a recurrent challenge for causal discovery and inference. In many situations, complex causal mechanisms only manifest themselves in extreme events, or take simpler forms in the extremes. Stimulated by data on extreme river flows and precipitation, we introduce a new causal discovery methodology for heavy-tailed variables that allows the effect of a known potential confounder to be almost entirely removed when the variables have comparable tails, and also decreases it sufficiently to enable correct causal inference when the confounder has a heavier tail. We also introduce a new parametric estimator for the existing causal tail coefficient and a permutation test. Simulations show that the methods work well and the ideas are applied to the motivating dataset.

Keywords Causation · Causal tail coefficient · Confounder · Extreme value statistics · Generalized Pareto distribution

1 Introduction

The field of causal inference has developed massively in recent decades (e.g., Pearl 2009; Peters et al. 2017), with much recent work on the detection of causality from observational data (e.g., Maathuis and Nandy 2016). Most of this literature concerns

✉ Olivier C. Pasche
olivier.pasche@unige.ch

Valérie Chavez-Demoulin
valerie.chavez@unil.ch

Anthony C. Davison
anthony.davison@epfl.ch

[1] Research Center for Statistics, University of Geneva, Geneva, Switzerland

[2] Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland

[3] Institute of Mathematics, EPFL, Lausanne 1015, Switzerland

central quantities such as expectations, but certain causal mechanisms manifest themselves only in rare events and/or may simplify in distribution tails. Standard methods of causal inference are ill-suited for such situations, and recent work has begun to link causality and extreme value theory. Examples are Gissibl and Klüppelberg (2018), who define recursive max-linear models on directed acyclic graphs, Klüppelberg and Krali (2021), who propose a scaling technique to determine the causal order of the variables in such graphs, Kiriliouk and Naveau (2020), who use multivariate generalized Pareto distributions to study probabilities of necessary and sufficient causation as defined in the counterfactual theory of Pearl, and Mhalla et al. (2020), who construct a causal inference method for tail quantities relying on Kolmogorov complexity of extreme conditional quantiles. See surveys by Naveau et al. (2020) on extreme event attribution and by Engelke and Ivanovs (2021) on the detection and modeling of sparse patterns in extremes.

Our work stems from that of Gnecco et al. (2021), who propose an estimator of the causal tail coefficient and an algorithm that, under mild conditions, consistently retrieves a causal order on an underlying graph even in the presence of hidden confounders. Such an order helps to exclude some causal structures, but does not provide evidence for the existence of a specific structure, as in general a given order is causal for several possible graphs; in particular, all orders are causal for the empty graph corresponding to absence of causality. Although it is asymptotically invariant to hidden confounders, this estimator can suffer from confounding in finite samples when inference on the direct relationship between two variables is needed, when these effects are too strong or when the confounders have heavier tails than the two variables.

This paper addresses a central challenge in causal inference: the presence of confounders. In theoretical development it is often assumed that all the relevant variables are observed and can be included in the model, but in practice one can rarely be sure of this. The available variables are often subject to external influences, observed or unobserved, that affect the variables of interest and can make it harder or even impossible to infer a correct causal relationship. Our goals are to mitigate the effect of a set of known confounders on an extremal causal analysis by treating them as covariates, and to present a permutation test for direct causality between the two observed variables. Our approach relaxes the assumption of Gnecco et al. (2021) that the confounders have the same tail index as the two main variables of interest, and thus encompasses a much broader range of situations, such as that in our application. Such a model enables causal discovery and inference for a greater variety of situations.

Our work was stimulated by average daily discharge data from 68 gauging stations along the Rhine and Aare catchments in Switzerland, see Fig. 1. The data were collected by the Swiss Federal Office for the Environment (hydrodaten.admin.ch), but were provided by the authors of Engelke and Ivanovs (2021), with some useful preliminary insights. We focus on the causal relationship between extreme discharges, for which precipitation is an obvious confounder, and use daily precipitation data from 105 meteorological stations, provided by the Swiss Federal Office of Meteorology and Climatology, MeteoSwiss (gate.meteoswiss.ch/idaweb). Unlike in our simulation experiments, we know neither the true tail properties of the discharges and precipitation nor the effect of the confounder. We

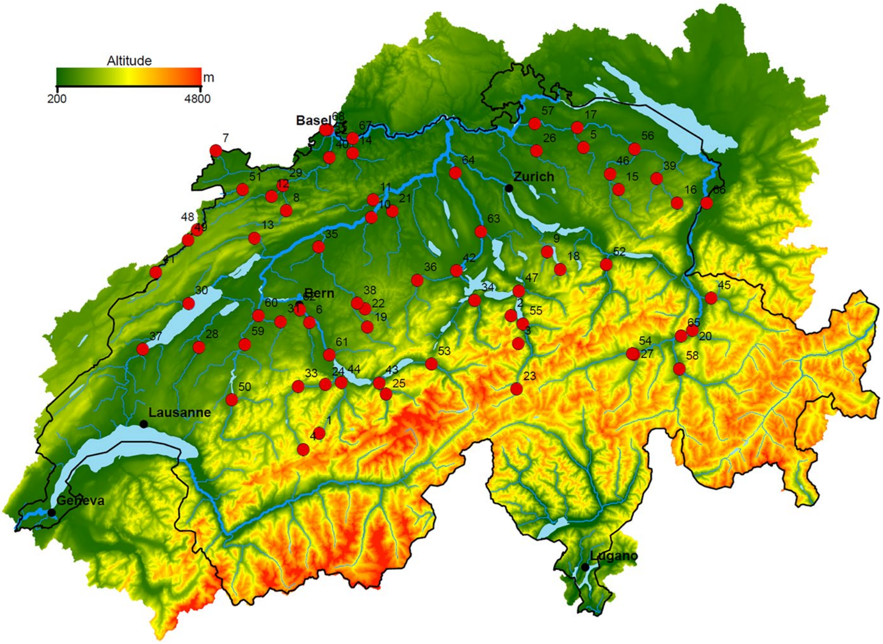

Fig. 1 Topographic map of Switzerland showing the 68 gauging stations (red dots) along the Rhine, the Aare and their tributaries. Water flows towards station 68. Adapted from Engelke and Ivanovs (2021)

use precipitation as a covariate in our test, allowing inference on the direct causal relationships between discharges for the majority of the station pairs, with at least 95% estimated confidence, which was impossible without our proposed approach.

The paper is organised as follows. Section 2 discusses the causal tail coefficient, its interpretation and its properties. Section 3 introduces a new parametric estimator for it based on generalized Pareto modelling of threshold excesses, which allows a known confounder to be used as a covariate. A simulation study in Section 4 underlines the strengths and limitations of the two estimators. Section 5 presents a permutation test intended to detect direct causality between two heavy-tailed variables, which is also assessed via simulation. Section 6 applies the methodology to the river discharges, and Section 7 gives a brief discussion.

2 Causal tail coefficient and its estimation

2.1 Existing work

We first give some basic notions needed to describe the setting in which causal relationships between random variables can be recovered.

Definition 1 A *linear structural causal model (LSCM)* over a set of random variables $X_1, \dots, X_p$ satisfies

$$X_j = \sum_{k \in \mathrm{pa}(j)} \beta_{jk} X_k + \varepsilon_j, \quad j \in V,$$

where $V := \{1, \dots, p\}$ is a set of nodes representing the corresponding random variables, $\mathrm{pa}(j) \subseteq V$ is the set of parents of j, $\beta_{jk} \in \mathbb{R} \setminus \{0\}$ is called the *causal weight* of node k on node j, and $\varepsilon_1, \dots, \varepsilon_p$ are jointly independent noise variables. We suppose that the *associated graph* $G = (V, E)$, in which the directed edge $(i,j) \in V \times V$ belongs to E if and only if $i \in \mathrm{pa}(j)$, is a directed acyclic graph (DAG).

In a DAG $G = (V, E)$, we say that $i \in V$ *is an ancestor of* $j \in V$ in G, if there exists a directed path from i to j. The set of the ancestors of j in G is denoted by $\mathrm{An}(j, G)$, and we define $\mathrm{an}(j, G) := \mathrm{An}(j, G) \setminus \{j\}$. In a LSCM over random variables $X_1, \dots, X_p$, with associated DAG $G = (V, E)$, we say that X_i *causes* X_j, if $i \in \mathrm{an}(j, G)$. We call X_i a *confounder* (or *common cause*) of X_j and X_k if there exist directed paths from i to j and from i to k in G that do not include k and j, respectively. We say that there is *no causal link* between X_i and X_j if $\mathrm{An}(i, G) \cap \mathrm{An}(j, G) = \emptyset$. For any $i,j \in V$ we let $\beta_{i \to j}$ denote the sum of the products of the causal weights along the distinct directed paths from vertex i to vertex j; we set $\beta_{j \to j} := 1$ and $\beta_{i \to j} := 0$ if $i \notin \mathrm{An}(j, G)$.

Let X_i and X_j be random variables from a LSCM with respective distributions F_i and F_j. The *causal (upper) tail coefficient* of a random variable X_i on another random variable X_j is defined as (Gnecco et al. 2021)

$$\Gamma_{ij} := \lim_{u \to 1^-} \mathbb{E}\{F_j(X_j) \mid F_i(X_i) > u\}, \tag{1}$$

if the limit exists. This coefficient lies between zero and one and captures the causal influence of X_i on X_j in their upper tails: if X_i has a linear causal effect on X_j, $\Gamma_{1,2}$ will be close to unity. The coefficient is asymmetric, as extremes of X_j need not lead to extremes of X_i, and in that case, Γ_{ji} will be appreciably smaller than Γ_{ij}. As Γ_{ij} only depends on the rescaled margins of the variables, it is invariant to monotone increasing marginal transformations.

If both tails are of interest, the causal tail coefficient can be generalized to capture the causal effects in both directions, by considering the *symmetric causal tail coefficient* of X_i on X_j, i.e.,

$$\Psi_{ij} := \lim_{u \to 1^-} \mathbb{E}\left[\rho\{F_j(X_j)\} \mid \rho\{F_i(X_i)\} > u\right]$$

if the limit exists, where $\rho : x \mapsto |2x - 1|$. As $F_i(X_i) \sim \mathrm{Unif}(0, 1)$,

$$\Psi_{ij} = \underbrace{\lim_{u \to 1^-} \frac{1}{2}\mathbb{E}\left[\rho\{F_j(X_j)\} \mid F_i(X_i) > u\right]}_{=:\Psi_{ij}^+} + \underbrace{\lim_{u \to 0^+} \frac{1}{2}\mathbb{E}\left[\rho\{F_j(X_j)\} \mid F_i(X_i) < u\right]}_{=:\Psi_{ij}^-}.$$

The interpretation and properties of Ψ_{ij} are similar to those of Γ_{ij}. The symmetric version captures the causal influence of X_i on X_j in both of their tails.

For simplicity we focus on Γ_{ij} in this paper, though all of our results and methods can be generalized to both tails by considering Ψ_{ij} instead, if the assumptions for the upper tails are also satisfied in the lower tails of the variables considered.

Before stating the theorem that describes how the underlying causal relationships in a set of random variables can be recovered, we define the concept of regular variation.

Definition 2 A positive measurable function f is said to be *regularly varying* with index $\alpha \in \mathbb{R}$, written $f \in \mathrm{RV}_\alpha$, if for all $c > 0$, $\lim_{x\to\infty} f(cx)/f(x) = c^\alpha$. If $f \in \mathrm{RV}_0$, then f is said to be *slowly varying*.

Definition 3 The random variable X_j is said to be *regularly varying* with index $\alpha > 0$, if, for some $\ell \in \mathrm{RV}_0$, $\mathbb{P}(X_j > x) \sim \ell(x)x^{-\alpha}$ as $x \to \infty$.

Independent regularly varying random variables $X_1, \dots, X_p$ are said to have *comparable upper tails* if there exist $c_1, \dots, c_p > 0$, $\alpha > 0$ and $\ell \in \mathrm{RV}_0$ such that, for each $j \in \{1, \dots, p\}$, $\mathbb{P}(X_j > x) \sim c_j\ell(x)x^{-\alpha}$ as $x \to \infty$.

The following theorem describes how the causal relationships underlying a set of random variables can be recovered from their causal tail coefficients.

Theorem 1 (Gnecco et al. 2021) *Let $X_1, \dots, X_p$ be random variables from a LSCM, with associated directed acyclic graph $G = (V, E)$ and suppose that*

(a) *the coefficients β_{jk} of the linear structural causal relationship $X_j = \sum_{k\in\mathrm{pa}(j,G)} \beta_{jk}X_k + \varepsilon_j$ are strictly positive for all $j \in V$ and $k \in \mathrm{pa}(j, G)$, and*

(b) *the real-valued noise variables $\varepsilon_1, \dots, \varepsilon_p$ are independent and regularly varying with comparable upper tails.*

Then the values of Γ_{ij} and Γ_{ji} allow one to distinguish between the different possible causal relationships between X_i and X_j summarized in Table 1.

Under the theorem's assumptions, the blank entries in Table 1 cannot occur. Theorem 1 is generalizable to the Ψ_{ij} variant of the coefficient and possibly negative β_{ij} values if the assumptions are also satisfied in the lower tails of the variables.

Gnecco et al. (2021) show that under the setup and assumptions of Theorem 1, the causal tail coefficient (1) for any distinct $i, j \in V$, and with $A_{ij} := \mathrm{An}(i, G) \cap \mathrm{An}(j, G)$, is

$$\Gamma_{ij} = \frac{1}{2} + \frac{1}{2}\frac{\sum_{h\in A_{ij}} \beta^\alpha_{h\to i}}{\sum_{h\in\mathrm{An}(i,G)} \beta^\alpha_{h\to i}}. \qquad (2)$$

Table 1 Equivalence of the possible values of Γ_{ij} and Γ_{ji} with the underlying causal relationship between X_i and X_j

	$\Gamma_{ji} = 1$	$\Gamma_{ji} \in (1/2, 1)$	$\Gamma_{ji} = 1/2$
$\Gamma_{ij} = 1$		X_i causes X_j	
$\Gamma_{ij} \in (1/2, 1)$	X_j causes X_i	common cause only	
$\Gamma_{ij} = 1/2$			no causal link

Without loss of generality we set $i = 1$ and $j = 2$ in what follows, and thus consider the causal effect of X_1 on X_2.

If $\left\{(X_{i,1}, X_{i,2})\right\}_{i=1}^{n}$ are independent replicates of (X_1, X_2), with the random variables X_i and X_j from the LSCM, then the *non-parametric estimator* of $\Gamma_{1,2}$ is defined to be

$$\hat{\Gamma}_{1,2} = \frac{1}{k} \sum_{i=1}^{n} \hat{F}_2(X_{i,2}) \mathbb{1}(X_{i,1} > X_{(n-k),1}) \tag{3}$$

for some $k \in \{1, \dots, n-1\}$, where $\mathbb{1}(\cdot)$ denotes the indicator function, $X_{(h),1}$ denotes the h^{th} order statistic and $\hat{F}_j$ is the empirical cumulative distribution function of X_j, i.e.,

$$\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{i,j} \leq x), \quad j = 1, 2.$$

This estimator is the empirical counterpart to (1), as $X_{(h),1} = \hat{F}_1^{\leftarrow}(h/n)$ is a quantile of the corresponding empirical distribution. The value of k controls the number of data pairs in the upper tail of X_1 that contribute to the estimator. Under the assumptions of Theorem 1 and a "very mild assumption that is satisfied by most univariate regularly varying distributions of interest", estimator (3) is consistent as $n \to \infty$, for a choice of k such that $k \to \infty$ and $k/n \to 0$ (Gnecco et al. 2021).

2.2 Practical limitations

A strength of the causal tail coefficient approach is its asymptotic robustness to hidden confounders. Studies of causation frequently presuppose that all the relevant variables have been observed, which is usually moot, but Theorem 1 holds even when some variables in the underlying LSCM are unobserved. This capacity to deal with confounders both when studying the causal relationship between two variables and when retrieving a causal order is not generally shared by other approaches in causal inference, as argued by Gnecco et al. (2021, Section 4.2), but the unobserved variables must satisfy a regular variation assumption that is hard to check and may be unrealistic. In practice, moreover, the tail behaviour of the confounders may differ from that of X_1 and X_2, violating assumption (b) of Theorem 1. In our motivating setting, for example, the tail of the confounder, precipitation, may not behave like the tails of the river discharges. This problem worsens when the confounder has a heavier tail than the variable of interest. Furthermore, distinguishing between different causal situations using empirical estimates may be difficult; an increase in the strength of the causal effect of a common confounder of X_1 and X_2 will increase $\Gamma_{1,2}$, making it harder to tell whether a high value of $\hat{\Gamma}_{1,2}$ indicates that $\Gamma_{1,2} = 1$ or that $\Gamma_{1,2} \lesssim 1$, as we shall see in Section 4.

The discussion above suggests that conditioning on the values of known confounders might be valuable. In the presence of a vector $\mathbf{H}$ of potential confounders we therefore define

$$\Gamma_{1,2|\mathbf{H}} := \lim_{u\to 1^-} \mathbb{E}_{(X_1,X_2,\mathbf{H})}\big\{F_2(X_2 \mid \mathbf{H}) \mid F_1(X_1 \mid \mathbf{H}) > u\big\}. \tag{4}$$

If there is no direct dependence of X_2 on X_1, then X_2 is independent of X_1 conditional on $\mathbf{H}$, so $\Gamma_{1,2|\mathbf{H}} = 1/2$, whereas $\Gamma_{1,2}$ lies in [1/2, 1) but might be close to unity. Thus $\Gamma_{1,2|\mathbf{H}} < \Gamma_{1,2}$ unless there are no confounders. If X_1 causes X_2, on the other hand, then $\Gamma_{1,2|\mathbf{H}} = \Gamma_{1,2} = 1$. In the presence of potential confounders, therefore, (4) seems preferable to $\Gamma_{1,2}$. The difficulty is that the estimation of (4) requires the modelling of the dependence of both X_1 and X_2 on $\mathbf{H}$. The first is more straightforward, because for large u only the upper tail of X_1 need be considered, whereas the second ostensibly requires a model for the entire distribution of X_2, and this may be complex. We compromise by fitting similar models to both variables, letting the upper tails alone vary with $\mathbf{H}$. As we shall see below, this can greatly improve estimation of the causal dependence structure relative to the original approach. Moreover fitting such a model should highlight simpler, potentially linear, structures in the tails, rather than more complex ones in the body of the data. This leads us to propose a peaks-over-threshold approach to estimating the conditional dependence of X_1 and X_2 on $\mathbf{H}$ (Section 3). Another useful tool, a reliable statistical test for direct causality, is discussed in Section 5.

3 Parametric tail causality and confounder dependence

3.1 Generalized Pareto causal tail coefficient

As mentioned above, we use the generalized Pareto distribution (GPD) to model the tails of our variables (Coles 2001, Chapter 4). For $j = 1, 2$, and under mild conditions on X_j, for a large enough threshold u_j large enough, we have

$$\mathbb{P}(X_j - u_j \le x \mid X_j > u_j) \approx G(x;\sigma_j, \xi_j) = 1 - \big(1 + \xi_j x/\sigma_j\big)_+^{-1/\xi_j}, \quad x > 0, \tag{5}$$

with a scale parameter $\sigma_j > 0$ and a shape parameter $\xi_j \in \mathbb{R}$:

- $\xi_j = 0$ corresponds to light-tailed distributions, and then X_j lies in the maximum domain of attraction of the Gumbel distribution;
- $\xi_j > 0$ corresponds to heavy-tailed distributions, and then X_j lies in the maximum domain of attraction of the Fréchet distribution; and
- $\xi_j < 0$ corresponds to distributions with bounded upper tails, and then X_j lies in the maximum domain of attraction of the (reverse) Weibull distribution.

Any random variable satisfying the assumptions of Theorem 1 satisfies (5), as a regularly varying random variable with index $\alpha > 0$ lies in the Fréchet maximum domain of attraction. If the threshold u_j is chosen to be the q quantile of X_j for some $q \in (0, 1)$, then we can write

$$\mathbb{P}(X_j \le x) \approx \left\{G(x-u_j;\sigma_j,\xi_j)(1-q)+q\right\}\mathbb{1}(x>u_j)+\mathbb{P}(X_j \le x)\mathbb{1}(x \le u_j),$$

and using the empirical distribution $\hat{F}(x)$ to estimate $\mathbb{P}(X_j \le x)$ and maximum likelihood estimation using the excesses of u_j to obtain $\hat{\sigma}_j$ and $\hat{\xi}_j$ yields a hybrid estimator of the distribution function $F_j(x)$ of X_j, i.e.,

$$\hat{F}_j(x;\hat{\sigma}_j,\hat{\xi}_j)=\hat{F}(x)\mathbb{1}(x \le u_j)+\left\{G(x-u_j;\hat{\sigma}_j,\hat{\xi}_j)(1-q)+q\right\}\mathbb{1}(x>u_j).$$

The choice of q involves a bias–variance trade-off: q should be chosen large enough for the tail to be well approximated by a GPD, thus reducing the bias, but small enough to have enough exceedances, thus reducing the variance of the estimator. Using hybrid estimators for F_1 and F_2 for an integer $k \in \{1,\dots,n-1\}$ yields the parametric *GPD causal tail coefficient* estimator for $\Gamma_{1,2}$,

$$\hat{\Gamma}^{\text{GPD}}_{1,2}=\frac{1}{k_g}\sum_{i=1}^{n}\hat{F}_2(X_{i,2};\hat{\sigma}_2,\hat{\xi}_2)\mathbb{1}\left\{\hat{F}_1(X_{i,1};\hat{\sigma}_1,\hat{\xi}_1)>1-k/n\right\}, \tag{6}$$

where $k_g := |\{i \in \{1,\dots,n\} : \hat{F}_1(X_{i,1};\hat{\sigma}_1,\hat{\xi}_1)>1-k/n\}|$. Unlike with the non-parametric estimator (3), the number of data pairs k_g used in (6) may not equal k, as it depends on the fit of $\hat{F}_1(X_{i,1};\hat{\sigma}_1,\hat{\xi}_1)$.

The GPD model can be extended to allow dependence on covariates of interest by expressing its parameters in the form $\theta(i)=h\{\boldsymbol{\gamma}^\top\mathbf{Z}(i)\}$, where θ denotes one or both of σ and ξ, h is an inverse link function, $\boldsymbol{\gamma}$ is a vector of parameters and $\mathbf{Z}(i)$ is the vector of explanatory variables on which the model might depend (Davison and Smith 1990).

We wish to reparametrise the model to reduce or remove the effect on $\Gamma_{1,2}$ of a vector of potential confounders $\mathbf{H}$ of X_1 and X_2. If $\mathbf{H}$ is part of the LSCM then under the setup in Section 2 it is straightforward to show that $\mathbf{H}$ affects the scale parameters of the GPD model that applies to X_1 and X_2 above high thresholds, but not their shapes, so we write

$$\sigma_j(i) := \sigma_j^0+\boldsymbol{\sigma}_j^{1\top}\mathbf{H}_i, \quad i=1,\dots,n,\ j=1,2, \tag{7}$$

where $\mathbf{H}_i$ is the replicate of $\mathbf{H}$ corresponding to the observations $(X_{i,1},X_{i,2})$ of (X_1,X_2).

This yields, for $k \in \{1,\dots,n-1\}$, the parametric $\mathbf{H}$-*conditional linear generalized Pareto distribution (LGPD) causal tail coefficient estimator*,

$$\hat{\Gamma}^{\text{GPD}}_{1,2|\mathbf{H}}=\frac{1}{k_l}\sum_{i=1}^{n}\hat{F}_2\{X_{i,2};\hat{\sigma}_2(i),\hat{\xi}_2\}\mathbb{1}\left[\hat{F}_1\{X_{i,1};\hat{\sigma}_1(i),\hat{\xi}_1\}>1-k/n\right]. \tag{8}$$

where $k_l := |\{i \in \{1,\dots,n\} : \hat{F}_1\{X_{i,1};\hat{\sigma}_1(i),\hat{\xi}_1\}>1-k/n\}|$. Estimation of σ_j^0, σ_j^1 and ξ_j is performed by maximum likelihood. In applications it is preferable to center and rescale each confounder in $\mathbf{H}$ componentwise to unit variance and zero mean, to avoid numerical issues. Although the confounder is here assumed to be part of the LSCM, this does not seem to be necessary in practice, as non-linear effects can be approximated linearly, especially in the tail region. We investigate the effect of varying the tail index in Section 4.2.

3.2 The positive linear scale issue

Linear modelling of the GPD scale parameter may not yield positive scale estimates $\hat{\sigma}_j(i) > 0$ for each $i = 1, \dots, n$ and $j = 1, 2$. The use of a nonlinear link function to ensure that the scale estimates were positive would not agree with the assumption of extremal linearity of the causal relationships, as the effect of $\mathbf{H}$ on the scale is also necessarily linear. We now describe two different solutions to this problem, which we compare by simulation in Section 4.

The first solution, *post-fit correction*, replaces $\hat{\sigma}_j(i)$ in (8) by $\max\{\hat{\sigma}_j(i), \epsilon\}$ for some arbitrary but small positive ϵ. The second solution, the *constrained approach*, applies the following linear constraints to the estimates when maximizing the likelihood

$$\sigma_j^0 + \boldsymbol{\sigma}_j^{1\top} \min_{i=1,\dots,n} \mathbf{H}_i > 0, \quad \sigma_j^0 + \boldsymbol{\sigma}_j^{1\top} \max_{i=1,\dots,n} \mathbf{H}_i > 0, \quad j = 1, 2, \tag{9}$$

where $\min_{i=1,\dots,n} \mathbf{H}_i$ and $\max_{i=1,\dots,n} \mathbf{H}_i$ represent the vectors of componentwise minima and maxima. When the data have a known distribution, box constraints can be used instead of (9). For example, in the case of a single confounder H and if X_1, X_2 and $X_h = H$ have t_ν distributions, then $\sigma_j^0 = u_j/\nu$ and $\sigma_j^1 = -\beta_{h\to i}/\nu$. Thus, if $\sigma_j(i) = \sigma_j^0 + \sigma_j^1 H_i > 0$ $(j = 1, 2; i = 1, \dots, n)$, then

$$-\frac{u_j}{\nu \max_{i=1,\dots,n} H_i} < \sigma_j^1 < -\frac{u_j}{\nu \min_{i=1,\dots,n} H_i}, \tag{10}$$

where the lower and upper bounds are needed for positive and negative H_i, respectively.

4 Simulation study

Here we perform a simulation study using the Student t, Pareto and log-normal noise distributions. The first two lie in the Fréchet maximum domain of attraction and are regularly varying with index $\alpha = 1/\xi > 0$. We write Pareto(a, α) for the Pareto model with scale parameter a and tail index α; recal that lower values of α indicate heavier tails. This distribution satisfies Definition 3 exactly, so one might expect Pareto data to show better behaviour than Student data. The log-normal distribution, LogN(μ, σ^2) lies in the maximum domain of attraction of the Gumbel distribution and is not regularly varying, but finite samples from it can appear to be heavy-tailed.

We focus on the behaviour of the causal tail coefficient estimators (3) and (8) between two variables X_1 and X_2 in their causal configurations, as shown in Fig. 2. As we study the estimators of causal effects of both X_1 on X_2 and of X_2 on X_1, we generated simulations only for the four causal cases, A, B, C and D. The LSCM causal weights $\beta_{2,1}$, β_{1h} and β_{2h} were chosen to equal 1.0, by default, for each existing edge in all four cases. Hence, in D, X_2 is caused by X_1 and the single confounder H with equal strength, even though H has another effect on X_2 through X_1.

Unless stated otherwise, each estimate is based on a random sample of $n = 10^6$ triples (X_1, X_2, H), of which $k = 2\lfloor n^{0.4} \rfloor = 502$ were chosen — Gnecco et al. (2021) found that the optimal fractional exponent of n for choosing k seems to lie between

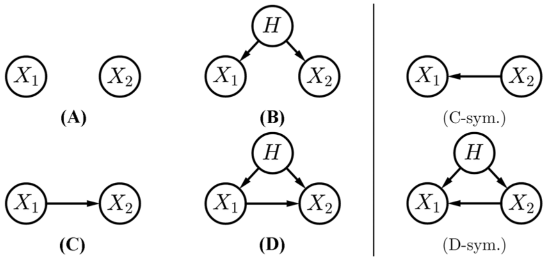

Fig. 2 The six possible causal configurations between X_1 and X_2 with a possible confounder H, separated into the four cases studied in the simulations, and the two omitted by symmetry

0.3 and 0.4. The factor 2 doubles the number of data pairs used in the estimator, thus decreasing its variability, but does not introduce much bias for such large values of n. The GPD-based estimators are based on the top $(1-q)n$ observations, where we take $q=0.9$, though only around k of the largest observations are used to estimate the coefficients Γ_{ij}. Setting $q=0.95$ yields similar results. One thousand independent replicates were generated for each of the four causal configurations and three distributions.

We present only the highlights of the study; the code and all the results are available from github.com/opasche/ExtremalCausalModelling.

4.1 Variables with comparable tails

Detailed results for variables with comparable tails may be found in Section S.1 of the Supplementary Material. In this case it is essentially always possible to infer the existence and direction of any causality between X_1 and X_2, based on the non-parametric or **H**-conditional LGPD estimators, (3) or (8), of $\Gamma_{1,2}$ and $\Gamma_{2,1}$ alone. When the causal effects of H on X_1 and X_2, i.e., β_{1h} and β_{2h}, are increased relative to the noise variance and any causal effect $\beta_{2,1}$ of X_1 on X_2, both $\Gamma_{1,2}$ and $\Gamma_{2,1}$ increase in configuration B, and $\Gamma_{2,1}$ increases in configurations C and D. This increase is larger with the non-parametric estimators of $\Gamma_{1,2}$ and $\Gamma_{2,1}$, which are biased upwards in these configurations. When the confounder has a high causal impact, inference based on the non-parametric estimator (3) for direct causal link between X_1 and X_2 can fail, as $\hat{\Gamma}_{1,2}, \hat{\Gamma}_{2,1} \approx 1$ and hence $|\hat{\Gamma}_{1,2}-\hat{\Gamma}_{2,1}| \approx 0$ in configurations B and D.

Use of the **H**-conditional LGPD estimator (8) greatly reduces the effect of H on the coefficient estimates in configurations B and D. For Pareto and log-normal data, the results are indistinguishable from those without the confounder, both in terms of location and variability, as if the effect of H had been entirely removed. The estimates based on Student data are also shifted to around the same values as in the corresponding confounder-free configurations, though their upper tails are marginally heavier. These few greater values remain appreciably lower than without H as a covariate. For configurations A and C, unlike for B and D, the estimator is almost unaffected by the addition of H as a covariate when it is not a confounder. This is also a useful property, as it could allow tests of whether a specific covariate is a confounder of two variables, based on changes to the estimated coefficients.

4.2 Confounder with a different tail

One generalisation allows the tail of the distribution of H to be heavier or lighter than those of X_1 and X_2. A lighter tail does not negatively affect whether the non-parametric and **H**-conditional LGPD estimators can infer a direct causal relationship between X_1 and X_2, as the tails of X_1 and X_2 then dominate. Figure 3 shows the sampling distributions of $\hat{\Gamma}_{1,2}$ and $\hat{\Gamma}_{2,1}$ for all four causal structures when the tail of H is heavier than those of X_1 and X_2. The true coefficient values are unknown, as assumption (b) of Theorem 1 is not satisfied, though the coefficient for comparable tails, (2), is shown for comparison.

When H has a heavier tail than X_1 and X_2, the non-parametric estimators $\hat{\Gamma}_{1,2}$ and $\hat{\Gamma}_{2,1}$ in configuration B and $\hat{\Gamma}_{2,1}$ in configuration D are shifted well towards unity. With an even heavier-tailed, Student t_2, distribution for H (not shown here), the Student results resemble those for the Pareto and log-normal distributions. In all these cases it becomes impossible to infer a direct causal relationship between X_1 and X_2, owing to the effect of the heavier confounder tail on the non-parametric estimators.

Figure 3 shows that in configurations B and D the non-parametric estimator is badly affected by the heavier tail of H. Figure 4, which displays the sample distributions of $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\mathrm{GPD}}_{2,1|H}$ with post-fit correction when the tail of H is heavier than those of X_1 and X_2, shows that the use of H as a covariate solves this problem: the estimates shift towards the coefficient values in the corresponding confounder-free cases, and consistently yield positive values of the difference of estimates $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H} - \hat{\Gamma}^{\mathrm{GPD}}_{2,1|H}$ for configuration D and differences centred at zero for configuration B; see also Section S.1 of the Supplementary Material. The estimates in configurations A and C, without the confounder causal effect, are barely changed by using H as a covariate.

Simulation results for $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\mathrm{GPD}}_{2,1|H}$ with the constrained fit are very similar to those for post-fit correction for the Pareto and log-normal distributions, but not for the Student distribution. Figure 5 shows the sample distribution of $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H}$ and $\hat{\Gamma}^{\mathrm{GPD}}_{2,1|H}$ with the constrained fit, for a heavier confounder tail. For the Student distribution, the confounder affects the estimator appreciably more for the constrained fit than for post-fit correction, compared to the non-parametric results. As the Student distribution is heavy in both tails, the lower constraint in (9) forces $\hat{\sigma}_j(i)$ $(j = 1, 2)$ to have an appreciably smaller slope, explaining this reduced effect. In configurations with a confounder, the absolute values of the constrained $\hat{\sigma}^1_j$ may be up ten times smaller than those for post-fit correction. With both approaches $\hat{\sigma}^1_j$ rarely differs greatly from zero for configurations without a confounder.

Both types of constraint yield very similar estimates for the Student distribution; see github.com/opasche/ExtremalCausalModelling.

To summarize, the simulations show that both the non-parametric estimator (3) and the **H**-conditional LGPD estimator (8) perform well when the theoretical assumptions are met and the influence of a hidden confounder is limited. When this influence grows, it becomes increasingly difficult to confidently infer the causal relationship between the variables using the non-parametric estimator, but the **H**-conditional LGPD estimator allows us to detect this relationship by reducing the effect of the confounding.

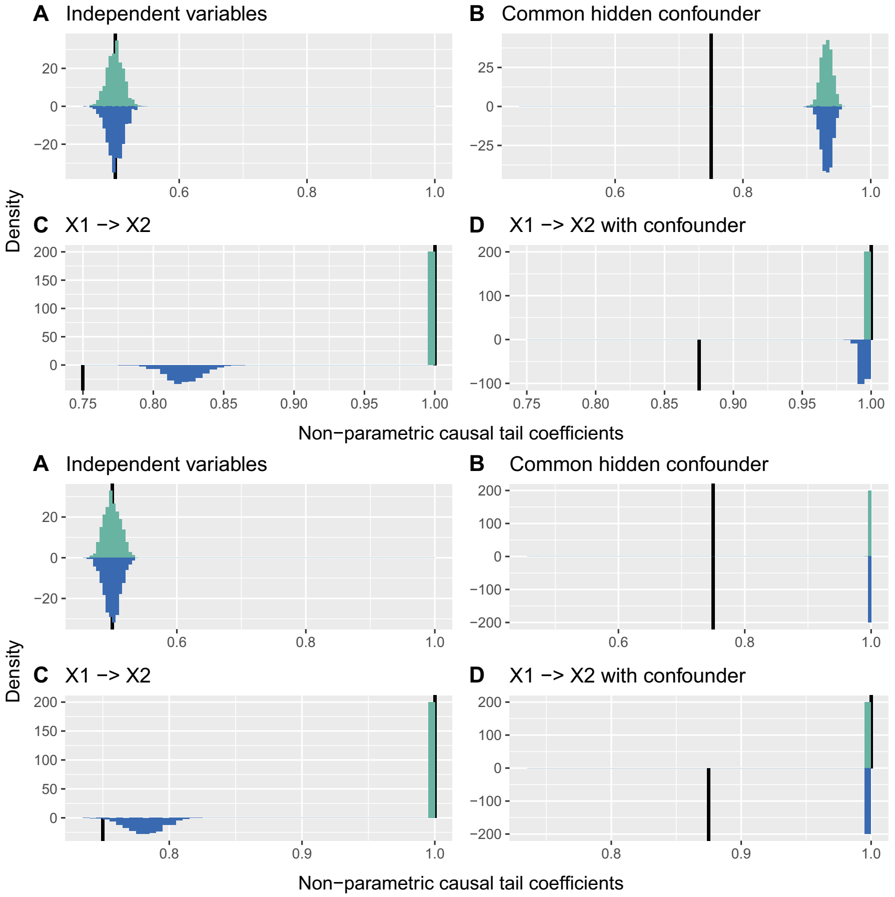

Fig. 3 Histograms of $\hat{\Gamma}_{1,2}$ (turquoise) and $\hat{\Gamma}_{2,1}$ (blue) for t_4-distributed ε_1 and ε_2, and t_3-distributed H (top four panels) and for LogN(0, 1)-distributed ε_1 and ε_2, and LogN(0, 1.5)-distributed H (bottom four panels). Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$ for comparable tails. The panels for Pareto(1, 3) distributed ε_1 and ε_2, and Pareto(1, 1.5) distributed H are very similar to the lower four panels

5 Testing for direct causality

5.1 Permutation test

In situations such as the causal analysis presented in Section 6, the distributions of the $\Gamma_{1,2}$ and $\Gamma_{2,1}$ estimators must be estimated to be used for inference. One way to obtain such distributions would be bootstrap resampling, but the extremal nature of the causal tail coefficient would require an unrealistically large sample size for its bootstrap distributions to be trustworthy, as these distributions tend to be too discrete in the extremes.

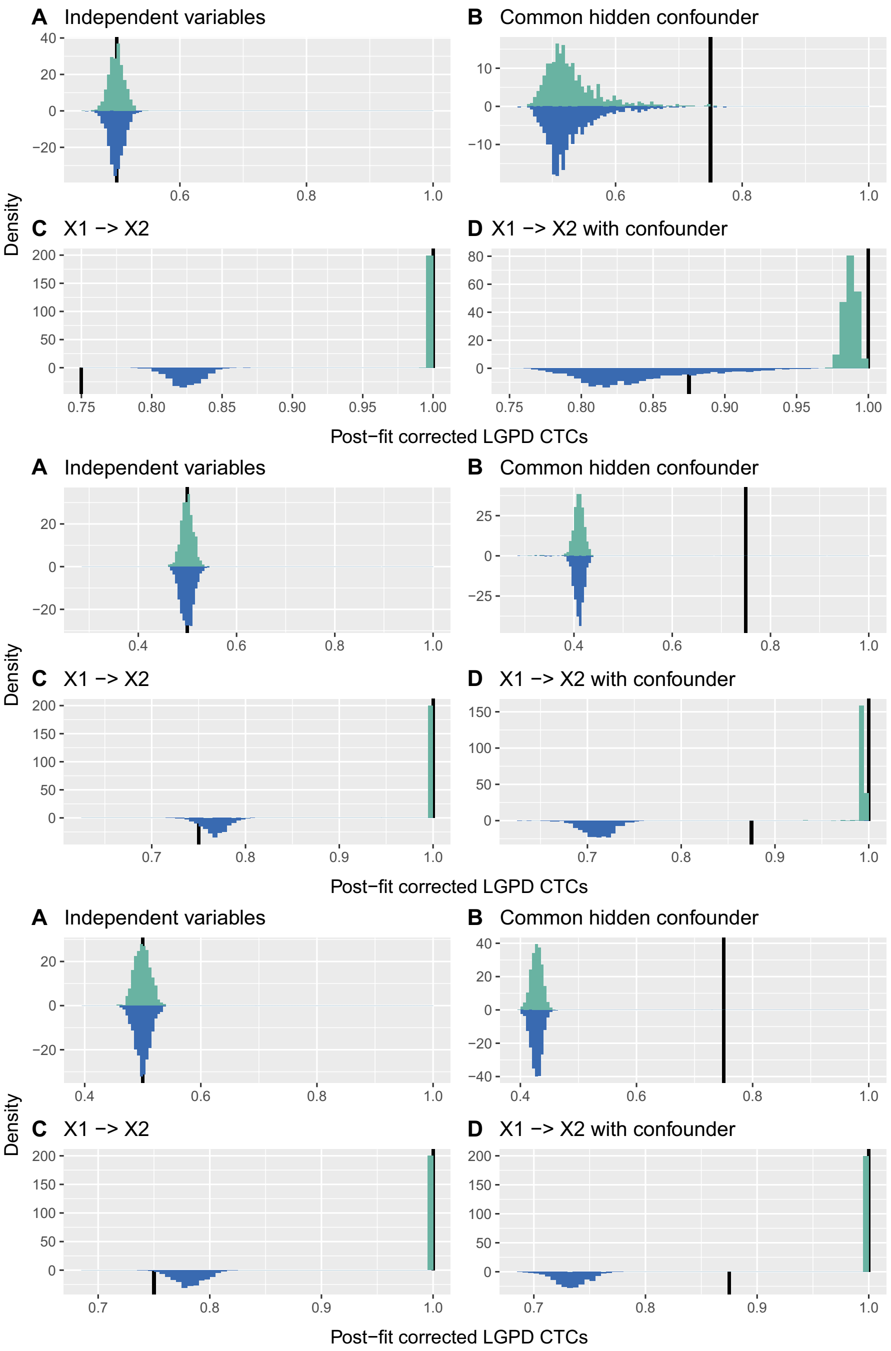

Fig. 4 Histograms of $\hat{\Gamma}^{\mathrm{GPD}}_{1,2|H}$ (turquoise) and $\hat{\Gamma}^{\mathrm{GPD}}_{2,1|H}$ (blue) with post-fit correction for t_4 distributed ε_1 and ε_2, and t_3 distributed H (top four panels), for Pareto(1, 3) distributed ε_1 and ε_2, and Pareto(1, 1.5) distributed H (middle four panels), and LogN(0, 1) distributed ε_1 and ε_2, and LogN(0, 1.5) distributed H (lower four panels). Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$ for comparable tails

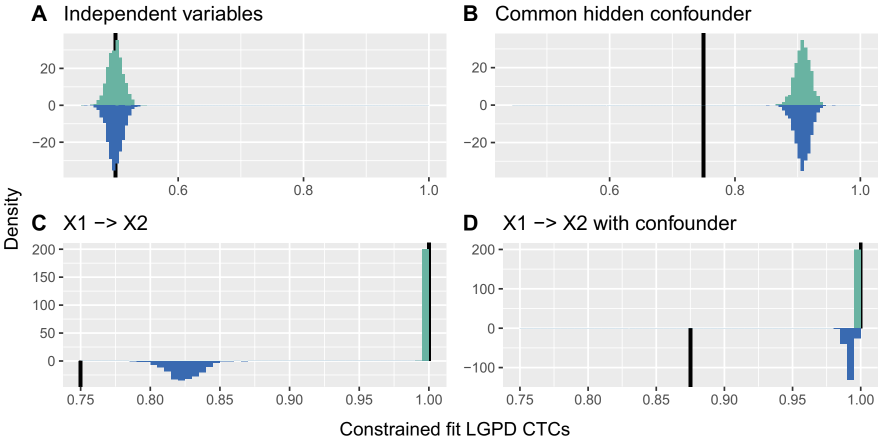

Fig. 5 Histograms of $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$ (turquoise) and $\hat{\Gamma}^{\text{GPD}}_{2,1|H}$ (blue) with constrained fit for t_4 distributed ε_1 and ε_2, and t_3 distributed H. Half-lines (black) indicate $\Gamma_{1,2}$ and $\Gamma_{2,1}$ for comparable tails

We therefore propose a permutation test (Davison and Hinkley 1997, Chapter 4) for direct causality between two observed variables, measuring the asymmetry in their direct causal relationship. Suppose we have a sample $\left\{(X_{i,1}, X_{i,2})\right\}_{i=1}^{n}$ from a LSCM and wish to test the null hypothesis of no direct causal relationship between X_1 and X_2, $H_0 : \beta_{2,1} = 0$, versus the alternative that X_1 causes X_2, $H_A : \beta_{2,1} > 0$. Our proposed procedure is as follows:

1. Rescale values $\tilde{X}_{i,j} = \tilde{F}_j(X_{i,j})$ $(i = 1, \dots, n,\ j = 1, 2)$, where known confounders can be used in the distribution estimator $\tilde{F}_j$, as for $\hat{\Gamma}^{\text{GPD}}_{1,2|H}$.
2. For $r = 1, \dots, R$, obtain $\tilde{X}^{(r)}_{i,1}$ and $\tilde{X}^{(r)}_{i,2}$ by randomly permuting the indices $j = 1, 2$ for each pair $(\tilde{X}_{i,1}, \tilde{X}_{i,2})$ $(i = 1, \dots, n)$.
3. Compute $\tilde{\Delta}_{1,2} = \tilde{\Gamma}_{1,2} - \tilde{\Gamma}_{2,1}$ on the transformed original data $\{(\tilde{X}_{i,1}, \tilde{X}_{i,2})\}_{i=1}^{n}$ and $\tilde{\Delta}^{*r}_{1,2} = \tilde{\Gamma}^{*r}_{1,2} - \tilde{\Gamma}^{*r}_{2,1}$ on their bootstrapped values $\{(\tilde{X}^{(r)}_{i,1}, \tilde{X}^{(r)}_{i,2})\}_{i=1}^{n}$ $(r = 1, \dots, R)$.
4. Obtain the Monte Carlo p-value, by comparing the value of the test statistic on the original rescaled data with the permutation distribution,

$$p_{\text{mc}} = \frac{1 + \#_r\{\tilde{\Delta}^{*r}_{1,2} \geq \tilde{\Delta}_{1,2}\}}{R + 1}.$$

If there are no asymmetric confounding effects on the two variables, i.e. $\beta_{1h} = \beta_{2h}$ in the case of a single confounder, then $\Delta_{1,2} := \Gamma_{1,2} - \Gamma_{2,1} = 0$ under H_0, whereas $\Delta_{1,2} > 0$ under H_A; see Eq. (2) and Theorem 1. This does not hold generally with asymmetric confounding. The direct causal relationship is symmetric under H_0, i.e., X_2 is as likely to take extreme values when X_1 is extreme as is X_1 when X_2 is extreme. If so, then permutations such as those performed in step 2. are equally likely, so $\tilde{\Delta}_{1,2}, \tilde{\Delta}^{*1}_{1,2}, \dots, \tilde{\Delta}^{*R}_{1,2}$ have a common distribution centered around zero, and p_{mc} will be uniformly distributed. Under the alternative, the direct causal relationship is “asymmetric”, as X_2 is more likely to be extreme when X_1 is extreme than conversely; then $\tilde{\Delta}_{1,2}$ is more likely to lie in the upper tail of $\tilde{\Delta}^{*1}_{1,2}, \ldots, \tilde{\Delta}^{*R}_{1,2}$. Thus the distribution of p_{mc} will become increasingly skewed towards zero as the causal strength of X_1 on X_2 increases.

If all asymmetric confounding effects are captured in $\tilde{F}_j$ by estimating the distribution conditionally, X_1 and X_2 have comparable tails and causal effects behave linearly in the extremes, then the proposed procedure should provide a reliable p-value for testing direct causality of X_1 on X_2.

5.2 Simulations

We used simulation from different data distributions and for different causal configurations involving X_1, X_2 and a potential confounder H to assess our proposed test. We used values of 0, 0.01, 0.05, 0.1, 0.2 for the causal strength $\beta_{2,1}$ of X_1 on X_2, with confounding effects both present and absent. Symmetric ($\beta_{1H} = \beta_{2H} = 1$) and asymmetric ($\beta_{1H} = 0.8$ and $\beta_{2H} = 1$, or $\beta_{1H} = 1$ and $\beta_{2H} = 0.8$) confounding effects were considered, and the noise variable were Pareto, Student t and log-normal. We generated $m = 10^3$ replicate samples of $n = 10^4$ independent triples $(X_{i,1}, X_{i,2}, H_i)$ for each causal configuration and noise distribution. The sample size n was chosen closer to practical orders of magnitude, compared to our large-sample study in Section 4. Three versions of the permutation test were performed for each sample, corresponding to the causal tail coefficient estimators discussed in Sections 2 and 3: the non-parametric (3), and **H**-conditional LGPD (8) with either post-fit correction or constrained fit. Each used $R = 10^3$ permutations and the estimator hyper-parameters were set to $k = 2\lfloor n^{0.4} \rfloor = 78$ and $q = 0.9$.

Figure 6 shows uniform QQ-plots of p_{mc} for the Pareto and Student distributions, in the case of heavier confounder tail, with symmetric effects. In the absence of confounding the test behaves as expected in both cases, and adding dependence on the independent H variable in the modelling through the parametric estimators has no visible effect on the distribution of p_{mc} compared to the non-parametric approach. For the Pareto distribution, the test has a power of almost 0.9 for a direct causal strength of 0.01, and it behaves perfectly for higher causal strengths. For the Student distribution, the test reaches a power of 0.3 for a direct causal strength of 0.05, of 0.7 for causal strength of 0.1 and of near 1.0 for a causal strength of 0.2.

When the confounding effects are added, the test based on the non-parametric estimator fails for the Pareto distribution, as most of the p_{mc} then lie outside the 95% confidence bands, indicating that the distribution of p_{mc} is highly non-uniform. This is corrected when the value of the confounder is taken into account using the parametric approaches, with power 0.9 for a direct causal strength of only one twentieth of the confounder’s marginal effects. In the Student case, p_{mc} seems to be close to uniformity in the absence of direct causality (the difference in tail shape is much greater in the Pareto case), but post-fit correction increases the power from below 0.2 to above 0.4 for a direct causal strength of one fifth of the confounder’s marginal effects. Similar conclusions to those of Section 4.2 about the constrained fit for

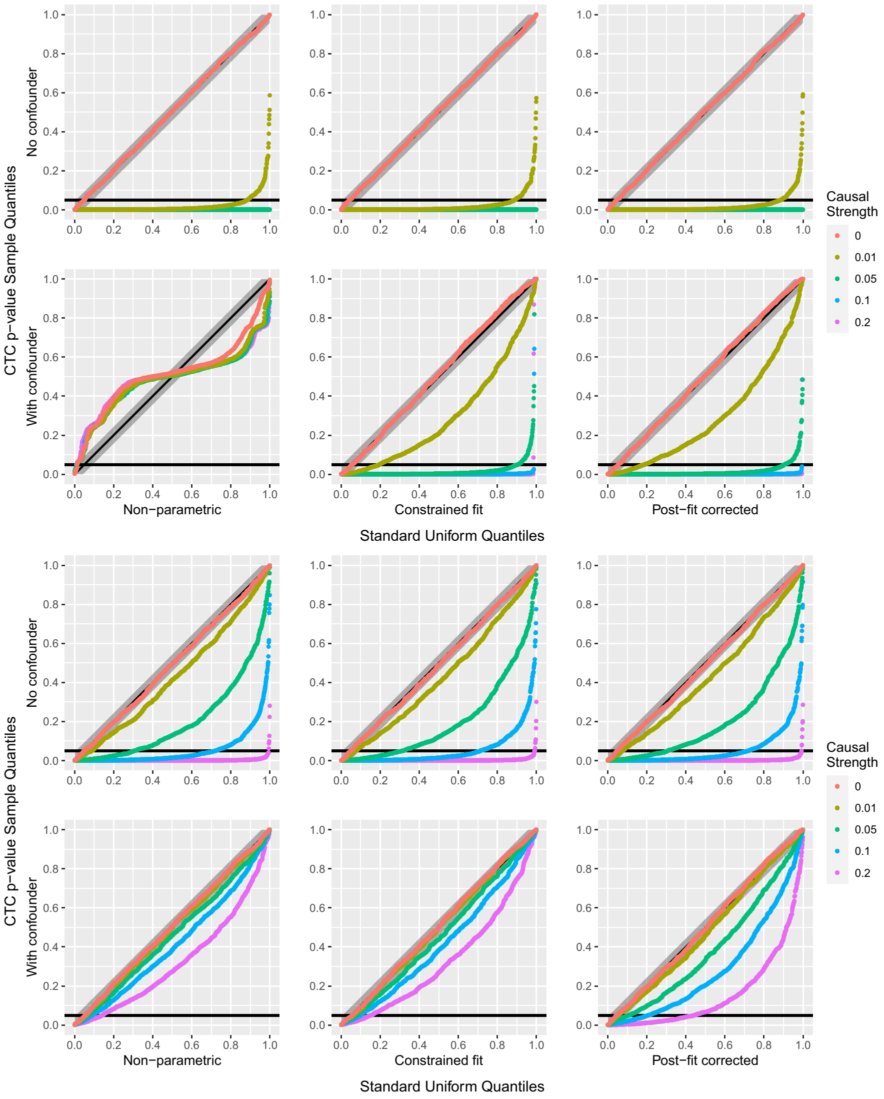

Fig. 6 Uniform QQ-plots of Monte Carlo p-values p_{mc}, with Kolmogorov–Smirnov confidence bands for different causal strengths $\beta_{2,1}$ (colors), the three estimators (columns) and optional symmetric confounding effects, $\beta_{1H} = \beta_{2H} = 1$ (rows). Top six panels: Pareto$(1, 2)$ distributed ε_1 and ε_2, and Pareto$(1, 1)$ distributed H. Bottom six panels: t_4 distributed ε_1 and ε_2, and t_3 distributed H

distributions with both tails heavy apply, as the constrained fit estimator is not significantly better than the non-parametric estimator compared to post-fit correction.

Figure 7 shows the uniform QQ-plots with asymmetric confounding effects for the Pareto distribution with comparable tails. Unlike in the corresponding symmetric case, the test here fails when using the non-parametric estimator owing to the asymmetry induced by the confounder, but both parametric approaches remove this

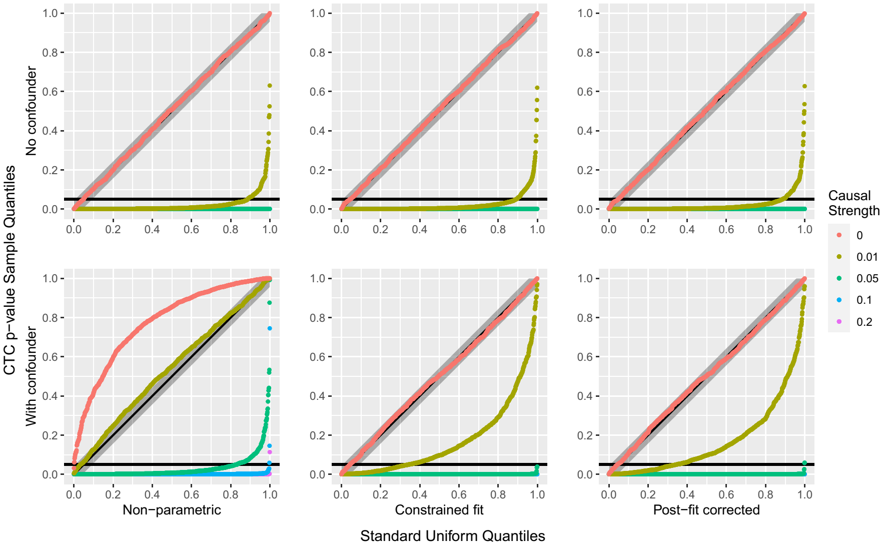

Fig. 7 QQ-plot of the p_{mc} estimates against the standard uniform distribution, with Kolmogorov–Smirnov confidence bands, for $\mathrm{Pareto}(1, 2)$ distributed ε_1, ε_2 and H, for different causal strengths $\beta_{2,1}$ (colors), the three estimators (columns) and optional asymmetric confounding effects, $\beta_{1H} = 0.8$, $\beta_{2H} = 1$ (rows)

unwanted effect by enough that p_{mc} nearly has a uniform distribution, with almost perfect power, for a causal strength of one sixteenth and one twentieth of the marginal confounding effects.

6 Application to Swiss rivers

We now illustrate how our method can discover direct causal relationships between the discharge extremes of pairs of river stations. This illustrates our method on a real example for which we know the 'ground truth' of extremal causality, but unlike in the simulations of Section 4, we cannot control and do not know the true tail behaviour of the station discharges and their potential confounders.

6.1 Data sources and additional collection

We use the average daily discharges between January 1913 and December 2014 at the 68 Swiss gauging stations shown in Fig. 1, and add daily precipitation data from 105 meteorological stations during the same period. Some additional information, such as the station elevation, catchment surface area and mean elevation, glaciation percent and coordinates, was collected from the Federal Office for the Environment's website. To reduce any seasonal effects due to unobserved confounders, we only consider data during June, July and August, as the more extreme observations happen during this period when mountain rivers are less likely to be frozen.

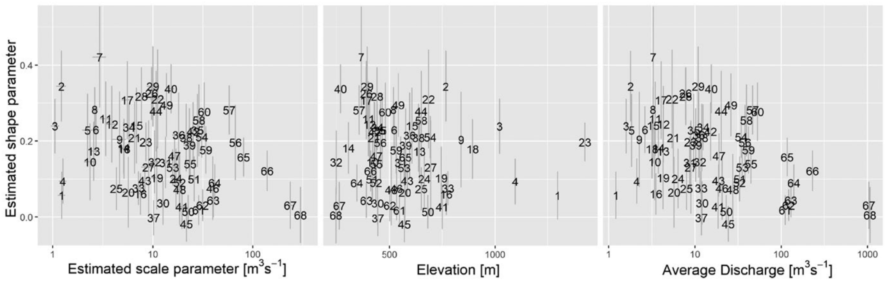

Fig. 8 Relation between shape parameter estimates, scale parameter estimates (log scale), station elevation and average discharge (log scale), with standard errors (±SE) shown as error bars

Temporal clustering is likely to appear for average daily discharge data but can be captured by considering the average catchment precipitation as a covariate in the model for the GPD scale parameter (7).

Figure 8 shows relationships between the estimates, station altitudes and average discharges. Altitude does not greatly affect the estimates, but the shape parameter estimates broadly decrease with increased average river discharge volume.

6.2 Choice of stations and comonotonicity

For the causal analysis, we consider pairs of stations with known direct causal relationships, and pairs with no direct causal relationship. Causal pairs are ordered by the flow of water, with one downstream of the other. The river volumes for the pairs should be as similar as possible, as our exploratory analysis indicated different tail behaviours for rivers with very different average discharges. There should also be enough confluences between the two stations, otherwise one would observe *comonotonicity*, i.e., almost perfect dependence, between their discharges. If there is comonotonicity between X_1 and X_2, then $F_1(X_{i,1}) \approx F_2(X_{i,2})$, for all $i = 1, \dots, n$, and it is impossible to know which variable causes which based on the data alone regardless of the approach, even if one is certain both of direct causality and of its direction. Confluences between the two stations reduce comonotonicity and make it possible to detect the direction of causality.

As we shall use precipitation as the confounding covariate, the stations must share likely meteorological effects and must lie in regions where precipitation data is available. Based on these criteria, we chose seven causal station pairs: (43,62), (42,63), (36,63), (24,61), (44,61), (22,38), (22,35), where the first station of each pair lies upstream from the second.

The non-causal station pairs were selected to have similar average volume and similar shape parameter estimates. Pairs with stations separated by long distances and pairs relatively close to each other were both considered. The 13 pairs selected are (30,45), (36,39), (42,34), (32,33), (62,63), (57,60), (13,14), (17,22), (12,21), (26,28), (27,31), (23,39), (23,35).

The choice of covariate for the causal pairs was the mean daily precipitation among the meteorological stations in the area and the catchment of the two stations.

The choice of covariate was less meaningful for the non-causal pairs with large separating distances, which have different meteorological conditions, so the average daily precipitation over the whole country was used. For the pair (42,34), which has the closest stations and local precipitation data available, the daily average in the local catchments was also considered. In the latter case, the pair will be highlighted with an asterisk to avoid confusion.

6.3 Causal analysis results

For each station pair, the permutation test for direct causality was performed using the non-parametric (3) and $\mathbf{H}$-conditional LGPD (8) estimators with post-fit correction or constraints, with $R = 10^4$ permutations and estimator hyper-parameters $k = 1.5\lfloor n^{0.4} \rfloor$ and $q = 0.9$. Table 2 shows the values of p_{mc}, the covariate shape estimate and its estimated extremal linear effects for the two stations, the latter estimated without constraints. The number of common observations for the pairs varies from 2024 to 8464, and k lies between 31 and 55. With precipitation covariates added, the number of common observations ranges from 1483 to 7820, and k lies between 27 and 54.

With the non-parametric approach for the causal stations, the absence of direct causality was rejected for four of the seven station pairs at significance level 5%, and for two of these four at level 2.5%. Adding daily precipitation as a covariate by either parametric approach decreases the p-values but two pairs remain non-significant; both lie in the same region and contain station 22.

With the non-parametric approach, the absence of direct causality was not rejected for ten of the 13 non-causal station pairs. Adding precipitation as a covariate with the two parametric approaches 'corrected' the p-value for another station. For the pair (42, 34) using local instead of global precipitation as a covariate gave a higher p-value.

We also considered using an exponential rather than a linear inverse-link function, i.e., taking $\log \sigma_j(i) = \sigma_j^0 + \sigma_j^1 H_i$ $(i = 1, \dots, n; j = 1, 2)$, to avoid any need for correction or constraints. The resulting p_{mc} values, also shown in Table 2, lead to the same conclusions as with the linear approaches.

Using the usual normal approximation, every $\hat{\sigma}_1^1$ is significantly positive for the causal pairs and 10 of the 14 estimates are positive for the non-causal pairs, with the highest confidence for the pair using local precipitation. Standard errors for $\hat{\sigma}_2^1$ are systematically larger than those for $\hat{\sigma}_1^1$ for the causal pairs, perhaps owing to the double causal effect of the covariate on the downstream station, both direct and indirect through the upstream station, as we do not observe this systematically for non-causal pairs. Consequently, the $\hat{\sigma}_1^1$ estimates are significantly positive for only four of the seven causal pairs, to be contrasted with 12 of the 14 estimates for the non-causal pairs. In particular, only the local precipitation effect is significant for the pair (42, 34).

We compare our results to two classical causal inference approaches appropriate to our problem. These are a non-Gaussian method for estimating causal linear structures based on results from independent component analysis, ICA-LiNGAM (Shimizu et al. 2006), and the PC algorithm, which retrieves the completed partially directed

Table 2 Permutation p-values p_{mc} for station pairs using the non-parametric approach (NP), the $\mathbf{H}$-conditional post-fit corrected (PFC) and constrained fit (CF) LGPD approaches, and an $\mathbf{H}$-conditional exponential inverse-link GPD approach (Exp). The shape estimate $\hat{\xi}_H$ for the precipitation covariate and the unconstrained scale slope estimates are also shown (with standard errors of at most 0.03 for the former and in parentheses for the latter)

Stations	Pair type	NP	PFC	CF	Exp	$\hat{\xi}_H$	$\hat{\sigma}_1^1$	$\hat{\sigma}_2^1$
43-62	causal	0.01	0.01	0.01	0.01	0.06	0.88(0.3)	1.91(1.3)
42-63	causal	0.03	0.02	0.02	0.04	0.06	6.49(1.1)	8.60(2.2)
36-63	causal	0.03	0.02	0.02	0.03	0.06	5.03(1.1)	7.25(2.8)
24-61	causal	0.06	0.01	0.01	0.00	–0.01	3.42(1.2)	–2.34(2.4)
44-61	causal	0.01	0.00	0.00	0.01	0.01	1.89(0.7)	–1.21(2.0)
22-38	causal	0.58	0.40	0.40	0.33	0.07	3.43(0.8)	8.00(2.0)
22-35	causal	0.22	0.17	0.17	0.10	0.03	3.43(0.9)	11.67(3.0)
30-45	non-caus.	0.56	0.47	0.47	0.46	0.01	1.01(0.4)	0.89(0.9)
36-39	non-caus.	0.80	0.70	0.70	0.69	0.01	4.61(1.1)	4.17(1.6)
42-34	non-caus.	0.23	0.04	0.04	0.10	0.01	5.97(1.2)	0.43(0.3)
42-34*	non-caus.	0.23	0.13	0.13	0.11	0.05	6.29(1.1)	0.66(0.3)
32-33	non-caus.	0.01	0.01	0.01	0.00	0.01	0.63(0.4)	1.00(0.3)
62-63	non-caus.	0.10	0.49	0.48	0.30	0.01	1.08(1.4)	7.67(2.1)
57-60	non-caus.	0.99	1.00	1.00	1.00	0.01	6.31(3.7)	5.23(1.8)
13-14	non-caus.	0.32	0.56	0.56	0.53	0.01	0.59(0.2)	1.19(0.3)
17-22	non-caus.	0.01	0.05	0.06	0.05	0.01	0.78(0.5)	2.18(0.7)
12-21	non-caus.	0.51	0.50	0.50	0.72	0.01	0.71(0.3)	1.33(0.4)
26-28	non-caus.	0.63	0.90	0.89	0.92	0.01	1.90(0.5)	1.63(0.4)
27-31	non-caus.	0.40	0.63	0.62	0.75	0.01	1.71(0.7)	2.91(1.1)
23-39	non-caus.	0.80	0.91	0.92	0.93	0.01	2.50(0.6)	4.27(1.5)
23-35	non-caus.	0.65	0.88	0.89	0.86	0.01	2.50(0.6)	6.66(1.7)

*Highlights the pair 42-34 that only uses the daily average of local precipitation, as opposed to 42-34 and other "non-caus." pairs that use the average precipitation over the whole country. This does NOT apply to "causal" pairs! (as they all use local precipitation averages)

acyclic graph by performing conditional independence tests on the variables. For the latter, we consider both the classic PC algorithm (Spirtes et al. 2000), which uses Gaussian conditional independence tests, and the Rank PC algorithm (Harris and Drton 2013), which uses rank-based Spearman correlation to perform the independence tests and thus is more robust to non-normal variables. The results for the ICA-LiNGAM method are presented in Table S.1 in the Supplementary Material, which shows the linear causal coefficients for the discharge station pairs estimated with the ICA-LiNGAM algorithm using either the station pair only (two variables) or the station pair and precipitation (three variables). Non-null values indicate significant causal effects. The upper-script arrows indicate the estimated direct causal direction between the station pair. Although in both cases of the two or three variables, ICA-LiNGAM retrieves all the correct causal pairs, with correct direction, all the non-causal pairs are indicated by non-null values as significantly causal. Both versions of the PC algorithm, once applied to our 21 pairs, provide existing direct causal links (without weights nor direction) between all the pairs of stations. Apparently both ICA-LiNGAM and PC methods are too eager to detect causality, unlike the tail coefficients. One explanation could be a set of unobserved confounders related to common global weather conditions triggering causal effects even between stations that are far apart. Extreme discharges depend more on local weather conditions, and particularly on heavy precipitation. Another explanation could be that causal effects are only linear in the tails, perhaps due to ground saturation by precipitation.

7 Discussion and conclusion

This paper addresses the reduction or removal of the unwanted effect of known confounders from the extremal causal analysis between two variables and the discovery of extremal causal relationships using a parametric estimator of the causal tail coefficient, based on generalized Pareto modelling, and a permutation test for direct causality. Both allow the use of known confounders as covariates.

In our simulation study, the new estimator removed the confounder's unwanted effect almost entirely for variables with comparable tails, and reduced its effect enough to allow correct causal inference on the direct causal relationship in the case of a confounder with a heavier tail. The permutation test was shown to provide reliable p-values when all asymmetric confounding effects are captured in the model.

When applied to Swiss river discharge data, our methodology allowed correct inference on the direct causal relationships between discharges for the majority of the chosen station pairs, and the parametric approach captured the confounding effect of precipitation.

In many real-life situations, statistically significant covariates need not correspond to causal effects. Peters et al. (2016) have proposed a methodology for causal discovery, for when data from different settings or regimes are observed. Their method constructs invariant causal regression or classification models that should still make accurate predictions under interventions on the covariates or a change of environment. Adapting this approach to our setting would lead to a better understanding of causality of extremes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10687-022-00456-4.

Acknowledgements We thank the referees and associate editor for their helpful remarks.

Funding Open access funding provided by University of Geneva. This work was supported by the Swiss National Science Foundation.

Data availability The data that support the findings of this study may be obtained from the Swiss Federal Office for the Environment (hydrodaten.admin.ch) and the Swiss Federal Office of Meteorology and Climatology, MeteoSwiss (gate.meteoswiss.ch/idaweb) but restrictions apply, as the data were used under licence for the current study and so are not publicly available. They are however available from the authors upon reasonable request and with permission of the Swiss federal offices.

Declarations

Competing interests The authors have no relevant competing interests to disclose.

References

Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer, London, (2001). https://doi.org/10.1007/978-1-4471-3675-0

Davison, A.C., Hinkley, D.V.: Bootstrap Methods and their Application. Cambridge University Press, New York, (1997). https://doi.org/10.1017/CBO9780511802843

Davison, A.C., Smith, R.L.: Models for exceedances over high thresholds (with discussion). Journal of the Royal Statistical Society, Series B **52**, 393–442 (1990). https://doi.org/10.1111/j.2517-6161.1990.tb01796.x

Engelke, S., Ivanovs, J.: Sparse structures for multivariate extremes. Annual Review of Statistics and its Application **8**, 241–270 (2021). https://doi.org/10.1146/annurev-statistics-040620-041554

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

Gnecco, N., Meinshausen, N., Peters, J., et al.: Causal discovery in heavy-tailed models. Annals of Statistics **49**(3), 1755–1778 (2021). https://doi.org/10.1214/20-AOS2021

Harris, N., Drton, M.: PC Algorithm for Nonparanormal Graphical Models. Journal of Machine Learning Research **14**, 3365–3383 (2013)

Kiriliouk, A., Naveau, P.: Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory. Annals of Applied Statistics **14**(3), 1342–1358 (2020). https://doi.org/10.1214/20-AOAS1355

Klüppelberg, C., Krali, M.: Estimating an extreme bayesian network via scalings. Journal of Multivariate Analysis **181**(104), 672 (2021). https://doi.org/10.1016/j.jmva.2020.104672

Maathuis, M.H., Nandy, P.: A review of some recent advances in causal inference. In: Bhlmann, P., Drineas, P., Kane, M., van der Laan, M.J. (eds.) Handbook of Big Data. Chapman and Hall (2016)

Mhalla, L., Chavez-Demoulin, V., Dupuis, D.: Causal mechanism of extreme river discharges in the upper Danube basin network. Applied Statistics **69**, 741–764 (2020). https://doi.org/10.1111/rssc.12415

Naveau, P., Hannart, A., Ribes, A.: Statistical methods for extreme event attribution in climate science. Annual Review of Statistics and Its Application **7**, 89–110 (2020). https://doi.org/10.1146/annurev-statistics-031219-041314

Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, NY, USA, 2nd edn

Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals (with Discussion). Journal of the Royal Statistical Society, Series B **78**(5), 947–1012 (2016). https://doi.org/10.1111/rssb.12167

Peters, J., Janzing, D., Schölkopf, B.: Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press, Cambridge, MA (2017)

Shimizu, S., Hoyer, P.O., Hyvärinen, A., et al.: A Linear Non-Gaussian Acyclic Model for Causal Discovery. Journal of Machine Learning Research **7**, 2003–2030 (2006)

Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. MIT press, Cambridge, MA, USA (2000)