Article

Extreme Treatment Effect: Extrapolating Dose-Response Function into Extreme Treatment Domain

Juraj Bodik [1,2]

1 Faculty of Business and Economics (HEC Lausanne), University of Lausanne, 1015 Lausanne, Switzerland; juraj.bodik@unil.ch
2 Expertise Center for Climate Extremes (ECCE), University of Lausanne, 1015 Lausanne, Switzerland

Abstract: The potential outcomes framework serves as a fundamental tool for quantifying causal effects. The average dose–response function $\mu(t)$ (also called the effect curve) is typically of interest when dealing with a continuous treatment variable (exposure). The focus of this work is to determine the impact of an extreme level of treatment, potentially beyond the range of observed values—that is, estimating $\mu(t)$ for very large t. Our approach is grounded in the field of statistics known as extreme value theory. We outline key assumptions for the identifiability of the extreme treatment effect. Additionally, we present a novel and consistent estimation procedure that can potentially reduce the dimension of the confounders to at most 3. This is a significant result since typically, the estimation of $\mu(t)$ is very challenging due to high-dimensional confounders. In practical applications, our framework proves valuable when assessing the effects of scenarios such as drug overdoses, extreme river discharges, or extremely high temperatures on a variable of interest.

1. Introduction

Quantifying causal effects is a fundamental problem in many diverse fields of research [1–4]. Some prevalent examples include the impact of smoking on developing cancer [5], the influence of education on increased wages [6], the effects of various meteorological factors on precipitation [7] or the effect of policy design on various economy factors [8].

The potential outcomes framework [9] has been the fundamental language to express the notion of the causal effect. The crux of this framework lies in acknowledging that, in any given scenario, multiple potential outcomes exist based on different interventions or exposures [10]. This perspective challenges researchers to consider not only the observed outcome but also the unobserved outcomes that could materialize under alternative conditions. The typical focus in causal inference is on the binary treatment variable (exposure). However, binary treatment is unable to differentiate between different levels of the treatment variable. This issue can be partially solved by assuming a continuous treatment. For example, Refs. [11,12] proposed an estimator based on local linear smoothing. Ref. [13] discussed the combination of parametric and non-parametric models for the effect curve estimation. Refs. [14,15] utilized marginal structural causal models framework. Refs. [16,17] applied neural networks for the effect estimation. However, typical methods that work with a continuous treatment variable are not well suited for the inference that goes beyond the observed range of the data.

In this paper, we are interested in the *extreme treatment effect*, that is, the quantity of interest is the effect of an extreme level of treatment, outside of the observed range. Consider the following example from medicine: Assume that the data of a study (either

randomized or observational) are available to us, with the health status (Y) of patients and the corresponding dose of a medicine administered (T). The data available only depict Y when $T \leq 20$ mg. What if then, we would like to know the change in Y when the dose is increased to $T = 25$ mg? Answering this inquiry is hard, since we have zero data to answer it (this might be considered unethical to give such a dose to a patient), and we must rely on strong unverifiable assumptions and extrapolation. Additionally, in the case of observational studies, high-dimensional confounders pose yet another significant challenge.

The connection between causal inference and extreme value theory has been receiving increasing interest. Refs. [18,19] analyze the Extreme Quantile Treatment Effect (EQTE) of a binary treatment on a continuous, heavy-tailed outcome. The paper authored by [20] develops a method to estimate the EQTE and the Extreme Average Treatment Effect (EATE) for continuous treatment. Ref. [21] developed a framework for Granger-type causality in extremes. Some other approaches for causal discovery using the extreme values include [22–25]. Ref. [26] propose graphical models in the context of extremes. Ref. [27] analyzed the the effect of climate change on weather extremes. Ref. [28] proposed a framework for extreme event propagation. Ref. [29] study probabilities of necessary and sufficient causation as defined in the counterfactual theory using the multivariate generalized Pareto distributions. We contribute to this growing literature and provide a theoretically well-founded approach for estimation and inference of the extremal treatment effect.

Recent advancements in machine learning research have spotlighted the extrapolation capabilities of different models [30–33]. For instance, 'engression', as proposed by [34], presents a framework that serves as an extrapolating alternative to regression-based neural networks. Similarly, Ref. [35] explored extrapolation of conditional expectations by assuming that the maximum derivative occurs within the observed range of support, and while these approaches are not inherently causal, they can be construed as such under certain assumptions. Despite achieving cutting-edge performance, these methods encounter two primary limitations: difficulty handling multiple confounders and reliance on often uninterpretable extrapolation assumptions. In contrast, our framework focuses on the causal aspect of the extrapolation and can handle a large number of confounders. This is achieved under weak assumptions commonly embraced in the extreme value theory. Moreover, our framework relies on strong yet more interpretable extrapolation assumptions; while our primary focus is on linear regression, our approach has the flexibility to integrate with various machine learning methodologies, potentially improving the overall performance. However, this integration may come at the expense of losing interpretability for certain assumptions.

As for the application in this work, we consider a dataset describing extreme precipitation and river discharge levels in Switzerland. A historical record indicates a maximum precipitation level near Zurich's meteo-station on 6 June 2002, reaching an extreme of $111 \frac{\text{mm}}{\text{m}^2}$. This event coincided with the river Reuss (near Zurich) nearly breaching its banks, causing damage to adjacent settlements. We focus on the following question: how would the river discharge alter if the precipitation on that day were to reach $120 \frac{\text{mm}}{\text{m}^2}$? Would the river breach its banks under such circumstances? We anticipate that the effect of precipitation on river discharge may vary between the body of the distribution and its tail. This anticipation stems from several factors: During periods of light to moderate precipitation, the ground absorbs a significant portion of the rainfall, reducing its contribution to the river flow. In contrast, during severe rainstorms, a larger proportion of the precipitation directly contributes to the river flow, potentially resulting in a more pronounced impact on discharge levels. Therefore, we expect to observe differing, potentially more severe, impacts of extreme precipitation on discharge levels compared to moderate events, highlighting the importance of understanding such dynamics across varying levels of precipitation intensity.

The structure of this paper is as follows: we introduce the notation and preliminaries on causal inference and extreme value theory in Section 2. In Section 3, we present the main assumptions along with some simple theoretical implications. In Section 4, we introduce a practical statistical methodology for estimating an extreme treatment effect from the data. Section 5 explains our methodology using a simple simulated example and discusses

simulation results. In Section 6, we explore the application of inferring the effect of extreme precipitation on river discharge levels.

This manuscript includes five appendices: Appendix A introduces a second real-world application regarding the compressive strength of concrete, which has been relocated to the appendix for the sake of brevity. Appendix B contains a detailed simulation study, exploring the methodology under various conditions, including (1) a varying dimension $dim(\mathbf{X})$, (2) comparison with classical methods from the literature, (3) a hidden confounder, (4) different dependence structures, and (5) varying dose–response functions. Appendix C contains a detailed inference process for the river application described in Section 6. Appendix D provides a more detailed explanation of the bootstrap algorithm used in the inference process and presents the theory behind the consistency result. Finally, proofs can be found in Appendix E.

2. Problem Statement, Notation and Preliminaries

Following [36], we define dose–response functions in the potential outcomes framework. We consider the triplet of $(\mathbf{X}, T, Y)$, where $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$, $T \in \mathcal{T} \subseteq \mathbb{R}$, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ denote the confounders, treatment, and response variable, respectively, in an observational causal study. We assume a continuous treatment setting, where $\mathcal{T} = (\tau_L, \tau_R)$ for some $\tau_L, \tau_R \in \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$. Here, $\tau_R \in \mathbb{R} \cup \{\infty\}$ is the right endpoint of the support of T. For simplicity, assume that the right endpoint of the support of $T \mid \mathbf{X} = \mathbf{x}$ is equal to τ_R for all $\mathbf{x} \in \mathcal{X}$. Let $Y(t)$ be a set of potential outcomes corresponding to the hypothetical world in which $T = t$ is set deterministically. The fundamental problem of causal inference arise, since in the real world, each individual can only receive one treatment level T, and we only observe the corresponding outcome $Y = Y(T)$.

We observe a random sample $\{\mathbf{X}_i, T_i, Y_i\}_{i=1}^n$ of size $n \in \mathbb{N}$. It follows from our setting that, given the observed covariates, the distribution of the potential outcome for one unit is assumed to be unaffected by the particular treatment assignment of another unit (Stable Unit Treatment Value Assumption). We utilize the letter H for a (possible) hidden confounder. We denote vectors by bold letters. For any pair of continuous random variables Z and Z', we denote its probability distribution function $P_Z(\cdot)$, density function $p_Z(\cdot)$ and a conditional density $p_{Z|Z'}(\cdot \mid \cdot)$.

The average dose–response function and patient-specific dose–response function are defined as

$$\mu(t) = \mathbb{E}[Y(t)], \qquad \mu_{\mathbf{x}}(t) = \mathbb{E}[Y(t) \mid \mathbf{X} = \mathbf{x}],$$

respectively. Although the term "dose" is typically associated with the medical domain, we adopt here the term dose–response learning in its more general setup: estimating the causal effect of a treatment on an outcome across different (continuous) levels of treatment. Our objective is to learn the behavior of $\mu(t)$ or $\mu_{\mathbf{x}}(t)$ for $t \approx \tau_R$.

For a pair of real functions f_1, f_2, we employ the following notation: $f_1(t) \sim f_2(t)$ for $t \to \tau_R$, if $\lim_{t \to \tau_R} \frac{f_1(t)}{f_2(t)} = 1$. The sequence of random variables approaches the same distribution as the sample size grows. In the remainder of the paper, we assume that $\mu(t), \mu_{\mathbf{x}}(t)$ are continuous on some neighborhood of τ_R for all $\mathbf{x}$.

2.1. Classical Assumptions

Two classical assumptions in the literature for identifying the average dose–response function are as follows:

- **Unconfoundedness**: Given the observed covariates, the distribution of treatment is independent of the potential outcome. Formally, we have $T \perp\!\!\!\perp Y(t) \mid \mathbf{X}, \forall t \in \mathcal{T}$, where $\perp\!\!\!\perp$ denotes the independence of random variables.
- **Positivity**: $p_{T|\mathbf{X}}(t \mid \mathbf{x}) > 0$ for all $t, \mathbf{x}$, where $p_{T|\mathbf{X}}$ represents the conditional density function of the treatment given the covariates.

Under these assumptions, [36] showed the identifiability of the dose–response function via

$$\mu_{\mathbf{x}}(t) = \mathbb{E}[Y|T=t, \mathbf{X}=\mathbf{x}], \quad \text{and} \quad \mu(t) = \mathbb{E}[\mu_{\mathbf{X}}(t)] = \mathbb{E}[\mathbb{E}[Y|T=t, \mathbf{X}=\mathbf{x}]], \tag{1}$$

where the inner expectation is taken over Y, and the outer expectation is taken over $\mathbf{X}$.

Even if we are not willing to rely on the unconfoundedness assumption, it may often still be of interest to estimate the function $t \to \mathbb{E}[\mathbb{E}[Y|T=t, \mathbf{X}=\mathbf{x}]]$ as an adjusted measure of association, defined purely in terms of observed data. It may be interpreted as the average value of Y in a population with exposure fixed at $T=t$, but it is otherwise characteristic of the study population with respect to $\mathbf{X}$ [11,12,37].

When the positivity assumption is violated, a different type of extrapolation arises [38], which is different from the one considered in this work. This scenario occurs when the distributional support of variable T varies across different levels of confounding variables $\mathbf{X}$. Various approaches have been devised to confront this challenge, including propensity thresholding [39].

Several algorithms were proposed to estimate the function $\mu(t)$ in the body of the distribution of T. State-of-the-art methods estimate $\mu(t)$ via $\sum_{i:T_i \approx t} w_i Y_i$ for appropriate weights w_i what serve to "erase" the confounding effect of $\mathbf{X}$ [11,40–43]. Typically, the estimation of $\mu(t)$ involves a two-step procedure [5,36,44]. In the first step, we model the distribution $T \mid \mathbf{X}$, also known as the propensity. In the second step, we model the distribution of $Y \mid T$, suitably adjusted by the propensity, with the aim of mitigating the confounding effect of $\mathbf{X}$.

Ref. [36] introduced a generalized propensity score (GPS) defined as $e(t, \mathbf{x}) := p_{T|\mathbf{X}}(t \mid \mathbf{x})$. One common approach is to model $p_{T|\mathbf{X}}$ using a Gaussian model. In binary treatment cases (when $\mathcal{T} = \{0, 1\}$), the propensity score is a probability denoted as $e(1, \mathbf{X}) = P(T = 1 \mid \mathbf{X})$ and is typically modeled using logistic regression. Subsequently, we define weights w_i as $w_i := \frac{1}{\hat{e}(T_i, \mathbf{X}_i)}$ or stabilized weights $w_i := \frac{\hat{p}_T(T_i)}{e(T_i, \mathbf{X}_i)}$, where we additionally model and estimate the marginal distribution of T, denoted as $\hat{p}_T$.

In a similar vein, Ref. [5] introduced the concept of a "uniquely parameterized propensity function assumption," which states that for every value of $\mathbf{X}$, there exists a unique finite-dimensional parameter $\theta \in \Theta$ such that $e(\cdot \mid \mathbf{X})$ depends on $\mathbf{X}$ only through $\theta(\mathbf{X})$. Since $\theta(\mathbf{X})$ contains all information about the confounding, we only model $\mathbb{E}[Y \mid T = t, \theta(\mathbf{X}) = \mathbf{s}]$ instead of $\mathbb{E}[Y \mid T = t, \mathbf{X} = \mathbf{x}]$ in Equation (1). In a vast majority of applications, $\theta(\mathbf{X})$ corresponds to the parameters of a normal distribution; to the best of our knowledge, there has been no exploration of the extreme value distribution in this particular context.

2.2. Extreme Value Theory

When dealing with extreme values, it is easy to introduce a strong selection bias. A naive approach for estimating $\mu(t)$ for large values of t might involve only considering observations where t exceeds a certain threshold, denoted as τ, and computing $\mu(t)$ using conventional techniques, while disregarding all values below this threshold. This is a typical approach of many classical algorithms, which estimate $\mu(t)$ by focusing solely on a local neighborhood of observations around t. However, this approach can introduce a significant selection bias. In its extreme manifestation, all observations where t exceeds τ might exclusively pertain to men, for instance. The selection bias arises if the effect of T on Y differs between men and women (see Figure 2 in Section 5.1 with $\tau = 3$). We employ the Extreme Value Theory technique known as peaks-over-threshold to tackle this issue.

Extreme value theory is a sub-field of statistics that explores techniques for extrapolating the behavior (distribution) of T beyond the observed values. A limiting theory posits that the tail of T can be well approximated by the Generalized Pareto Distribution (GPD), as detailed in the following explanation.

Consider a sequence $(T_i)_{i \geq 1}$ of independent and identically distributed (iid) random variables with a common distribution F, and $M_n = \max_{i=1,...,n} T_i$ represents the running

maximum. It is well known [45] that if there exists a non-degenerate distribution G such that $\frac{M_n - b_n}{a_n} \xrightarrow{D} G$ as $n \to \infty$ for some sequences of constants $\{a_n, b_n\}_{n=1}^{\infty} \in \mathbb{R}_+^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}}$, then G falls within the Generalized Extreme Value (GEV) distribution family. This can equivalently be expressed using the following definition:

Definition 1 ([46]). *The distribution F is in the max domain of attraction of a generalized extreme value distribution (notation $F \in MDA(\gamma)$) if there exist $\gamma \in \mathbb{R}$ and sequences of constants $a_n > 0, b_n \in \mathbb{R}, n = 1, 2, \ldots$ such that $\lim_{n \to \infty} F^n(a_n x + b_n) = exp(-(1 + \gamma x)^{-1/\gamma})$ for all x satisfying $1 + \gamma x > 0$. In case $\gamma = 0$, the right side is interpreted as $exp(-e^{-x})$. The parameter γ is called the extreme value index (or shape index).*

This condition is mild as it is satisfied for most standard distributions, for example, the normal, Student-t and beta distributions. The following crucial theorem states that the tail of T can be well approximated by GPD if the distribution of T belongs to $MDA(\gamma)$.

Theorem 1 (Theorem 4.1 in [47]). *Let $T \sim F \in MDA(\gamma)$. Then, for large $\tau \approx \tau_R$, there exist $\sigma > 0, \gamma \in \mathbb{R}$ such that the distribution of $T - \tau \mid T > \tau$ is approximately $GPD(0, \sigma, \gamma)$.*

GPD distribution has three parameters, namely a location $\tau \in \mathbb{R}$, scale $\sigma > 0$ and a shape $\gamma \in \mathbb{R}$. Its distribution function takes the following form:

$$H(x) = \begin{cases} 1 - \left(1 + \gamma \frac{x - \tau}{\sigma}\right)^{-1/\gamma}, \gamma \neq 0, \\ 1 - \exp\left(-\frac{x - \tau}{\sigma}\right), \gamma = 0, \end{cases}$$

defined on the support $[\tau - \sigma/\gamma, \infty), (-\infty, \infty), (-\infty, \tau - \sigma/\gamma]$ for cases $\gamma < 0, \gamma = 0, \gamma > 0$, respectively. Cases when $\gamma > 0, \gamma = 0$, and $\gamma < 0$ correspond to the Fréchet, Gumbel, and Weibull distributions, respectively, [45,48].

Note that when the distribution of $T - \tau$ given $T > \tau$ follows a GPD with parameters $0, \sigma$, and γ, an equivalent assertion can be made that T given $T > \tau$ follows a GPD with parameters τ, σ, and γ. We denote $\theta = (\tau, \sigma, \gamma)^{\top}$.

Assumption 1. *We assume that the distributions T and $T \mid X$ are in the max domain of attraction of a generalized extreme value distribution.*

3. Our Tail Framework

We aim to model the effect of a treatment variable T in the context of extreme values of T. However, it is essential to approach the term 'extreme' with caution, considering the discrepancy between real-world implications and the interpretations within our model. Take, for instance, if T represents a drug dose in milligrams; our model operates under the assumption that T tends toward τ_R, which can be potentially larger than several kilograms. While this mathematical abstraction lacks practical significance—given that administering several kilograms of a drug is physically implausible—the model does include values of T that are arbitrarily large. Of course, we do not claim that our model performs well when T equals several kilograms but only that it performs well for T in the 'reasonable neighborhood' of the observed values.

3.1. Assumptions

We are not aiming to estimate the complete $\mu(t)$ but rather only its values for large t. Therefore, we can relax the classical assumptions for the identification of $\mu(t)$; what we specifically need is their tail counterparts.

Assumption 2 (Unconfoundedness in tail). *For all $x \in \mathcal{X}$, it holds that*

$$\mathbb{E}[Y(t) \mid X = x] \sim \mathbb{E}[Y \mid X = x, T = t] \qquad as \; t \to \tau_R. \qquad \text{(Unconfoundedness in tail)}$$

We always assume the existence of the expected values.

Rather than simply writing $t \to \tau_R$, we frequently opt for the notation $t(\mathbf{x}) \to \tau_R$ to emphasize its dependence on the random variable $\mathbf{X}$. Note that Assumption 2 is strictly less restrictive than the Unconfoundedness assumption introduced in Section 2.1.

Remark 1. *To provide some intuition regarding the permissiveness of Assumption 2, we rephrase our framework in the language of structural causal models (SCM, [49]). Assume that the data-generating process of the output Y is as follows:*

$$Y = f_Y(T, \mathbf{X}, H, \varepsilon), \quad \varepsilon \perp\!\!\!\perp (T, \mathbf{X}, H).$$

Here, H represents a (possible) latent confounder of T and Y. Then, the dose–response function has a form $\mu(t) = \mathbb{E}[f_Y(t, \mathbf{X}, H, \varepsilon)]$ where the expectation is taken with respect to $(\mathbf{X}, H, \varepsilon)^\top$.

Assumption 2 can be rephrased as follows: there exists a function $\tilde{f}_Y$ such that

$$f_Y(t, \mathbf{x}, h, e) \sim \tilde{f}_Y(t, \mathbf{x}, e) \text{ as } t \to \tau_R, \qquad \text{(Unconfoundedness in tail in SCM)}$$

for all admissible values of $\mathbf{x}, h, e$. This assumption is valid for example in additive models; that is, when $f_Y(t, \mathbf{x}, h, e) = \tilde{f}_Y(t, \mathbf{x}, e) + g(h)$ for some functions $\tilde{f}, g$.

Additionally, we restate the positivity assumption in the context of its tail counterpart.

Assumption 3 (Positivity in tail). $p_{T|\mathbf{X}}(t \mid \mathbf{x}) > 0$ *for all $\mathbf{x}$ and all $t > t_0$ for some $t_0 \in \mathcal{T}$, where $p_{T|\mathbf{X}}$ represents the conditional density function of the treatment given the covariates.*

Note that this assumption is weaker than Assumption 1.

3.2. Adjusting Only for $\theta(\mathbf{X})$

The following lemma serves as a tail counterpart of an identifiability for the classical framework. It states that, under Assumptions 2 and 3, the tail of the dose–response function is identifiable from the observational distribution via the propensity function $\pi_0(t, \mathbf{x})$.

Lemma 1 (Identifiability). *Under Assumptions 2 and 3, it holds that*

$$\mu(t) \sim \mathbb{E}\{\pi_0(T, \mathbf{X})Y \mid T = t\}, \quad \text{as } t \to \tau_R, \tag{2}$$

where $\pi_0(t, x) := \frac{p_T(t)}{p_{T|\mathbf{X}}(t|x)}$ is the (stabilized) propensity function.

Recall that the distribution of $T \mid \mathbf{X} = \mathbf{x}$, conditioned on $T > \tau(\mathbf{x})$ for large $\tau(\mathbf{x}) \approx \tau_R$, is approximately GPD with parameters $\theta(\mathbf{x}) = (\tau(\mathbf{x}), \sigma(\mathbf{x}), \gamma(\mathbf{x}))$. The following result suggests that instead of conditioning on (potentially high-dimensional) covariates $\mathbf{X}$, we only need to condition on $\theta(\mathbf{X})$.

Lemma 2. *Under Assumptions 1 and 2, for all $\mathbf{s}$ in the support of $\theta(\mathbf{X})$ holds*

$$\mathbb{E}[Y(t) \mid \theta(\mathbf{X}) = \mathbf{s}] \sim \mathbb{E}[Y \mid T = t, \theta(\mathbf{X}) = \mathbf{s}] \quad \text{for } t \to \tau_R. \tag{3}$$

Hence,

$$\mu(t) \sim \int \mathbb{E}[Y \mid T = t, \theta(\mathbf{x})] p_{\theta(\mathbf{X})}(\mathbf{x}) d\mathbf{x} \quad \text{for } t \to \tau_R.$$

Lemma 2 suggests that it is sufficient to condition only on $\theta(\mathbf{X})$ rather than on $\mathbf{X}$. This finding is pivotal for dimension reduction, effectively reducing the dimension from $dim(\mathbf{X})$ to at most 3. Nonetheless, this is merely a limiting result, and it introduces an approximation error into the GPD approximation for finite samples.

3.3. Model for the Conditional Expectation of Y Given a T

Under Assumption 2, modeling $\mu_{\mathbf{X}}$ reduces to a statistical modeling of $\mathbb{E}[Y \mid T, \mathbf{X}]$. Furthermore, under Assumptions 1 and 2, it reduces to modeling $\mathbb{E}[Y \mid T, \theta(\mathbf{X})]$. In principle, a wide range of models can be considered, ranging from simple linear models to non-parametric neural networks. The principle of Occam's razor suggests that, especially when extrapolating beyond the range of observed values, simpler models often prove to be the most effective choices [50]. The extrapolation capabilities of various models have recently garnered attention in machine learning research. Ref. [34] introduced the 'engression' framework as an extrapolating counterpart to regression-based neural networks, and while we build our framework under a linear model for simplicity, it is possible to utilize different models, such as the engression-based ones. Figure 1 illustrates the extrapolating properties of various commonly used models.
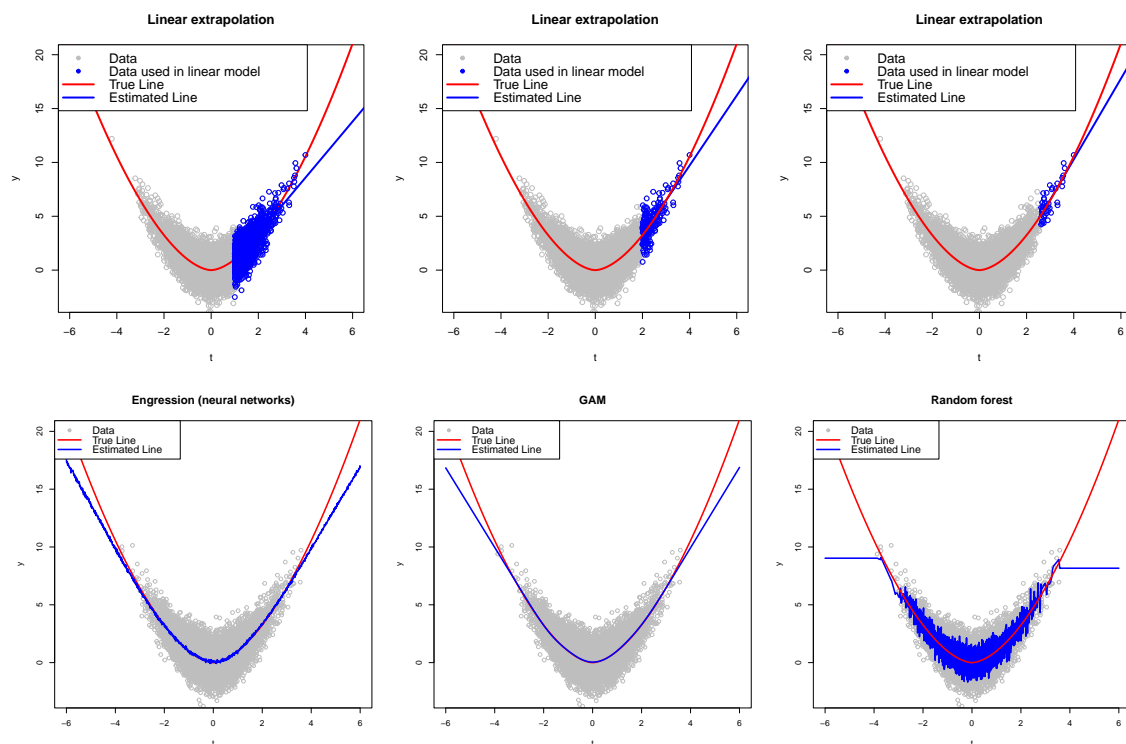
Figure 1. Extrapolation of $\mathbb{E}[Y \mid T = t]$ under various models (without confounding). The upper three figures illustrate a first-order extrapolation approach, employing a linear fit at the boundary of the support of T. In contrast, the lower three figures depict estimations generated by distinct models: the first utilizes a pre-additive noise model parameterized by neural networks [34], the second employs smoothing splines [51], and the third utilizes a random forest approach [52].

A straightforward approach to model $\mathbb{E}[Y \mid T = t, \mathbf{X} = \mathbf{x}]$ under the assumption of linearity-in-the-tail would be assuming an existence of functions $\tilde{\alpha}, \tilde{\beta}$ such that

$$\mathbb{E}[Y \mid T = t, \mathbf{X} = \mathbf{x}] \sim \tilde{\alpha}(\mathbf{x}) + \tilde{\beta}(\mathbf{x})t, \quad \text{as } t \to \tau_R. \tag{4}$$

Following the notation in Remark 1, this corresponds to assuming

$$f_Y(t, \mathbf{x}, h, e) \sim \tilde{\alpha}(\mathbf{x}) + \tilde{\beta}(\mathbf{x})t \text{ as } t \to \tau_R,$$

for all admissible values of h, e. This assumption is valid for example in additive models where $f_Y(t, \mathbf{x}, h, e) = \tilde{\alpha}(\mathbf{x}) + \tilde{\beta}(\mathbf{x})t + g(\mathbf{x}, h, e)$ for some function g. However, using the result from Lemma 2, it is sufficient to condition only on $\theta(\mathbf{X})$ instead of potentially high-dimensional $\mathbf{X}$. Therefore, we introduce the following model assumption:

Assumption 4 (Conditional linearity of tail). *There exist functions α and β such that for all s in the support of $\theta(X)$, the following holds:*

$$\mathbb{E}[Y \mid T = t, \theta(X) = s] \sim \alpha(s) + \beta(s)t, \quad as\ t \to \tau_R. \tag{5}$$

Such an assumption was explored in various contexts (typically where $\theta(X)$ represents parameters of a normal distribution, see [5] or Section 2.2.1 in [44]; to the best of our knowledge, the extreme case was not yet explored). We can construct our inference method by estimating α and β using various machine-learning methodologies. This is discussed in Section 4.

4. Inference and Estimation

Let $(\mathbf{x}_i, t_i, y_i)_{i=1}^n$ be the observed data. In the following, we propose a methodology for the estimation of $\mu(t)$ for $t \approx \tau_R$ under Assumptions 1, 2, and 4.

Consider the following two-step procedure. In the first step, we approximate the tail of $T \mid X$ using GPD (that is, we estimate the location, scale, and shape parameters $\theta(X) = (\tau(X), \sigma(X), \xi(X))$). In the second step, we estimate the expectation of Y given a large T conditional on the estimated GPD parameters $\hat{\theta}(X)$.

1. Estimate $\theta(\mathbf{x})$:
    *   Choose $q \in (0,1)$.
    *   Estimate the covariant-dependent threshold $\tau(\mathbf{x})$ using a quantile regression, that is, estimate q-quantile of $T \mid X = \mathbf{x}$.
    *   From now on, restrict our inference on the observations from $S := \{i : t_i > \hat{\tau}(\mathbf{x}_i)\}$.
    *   Estimate $\theta(\mathbf{x})$ in the tail model, that is, estimate (σ, ξ) from the data points in S in the model, where

$$T \mid T > \hat{\tau}(\mathbf{x}), X = \mathbf{x} \sim GPD(\hat{\tau}(\mathbf{x}), \sigma(\mathbf{x}), \xi(\mathbf{x})).$$

2. Estimate $\mu(t)$ or $\mu_{\mathbf{x}^\star}(t)$ using $\hat{\theta}(\mathbf{x})$:
    *   Estimate α, β in model (5) from the data points in S (that is, we only consider $t > \hat{\tau}(\mathbf{x})$).
    *   Return $\hat{\mu}(t) := \frac{1}{n} \sum_{i=1}^n \hat{\alpha}[\hat{\theta}(\mathbf{x}_i)] + \hat{\beta}[\hat{\theta}(\mathbf{x}_i)]t$ or $\hat{\mu}_{\mathbf{x}^\star}(t) := \hat{\alpha}[\hat{\theta}(\mathbf{x}^\star)] + \hat{\beta}[\hat{\theta}(\mathbf{x}^\star)]t$.

The first step is a very standard procedure in the extreme-value literature called 'peak-over-threshold' [47]; it is standard to assume a constant shape parameter $\gamma(\mathbf{x}) \equiv \gamma \in \mathbb{R}$ since in practice, it is untypical for the shape parameter to change with covariates [53,54]. For the estimation of $\tau(\mathbf{x}), \sigma(\mathbf{x}), \alpha(\mathbf{x}), \beta(\mathbf{x})$, we use either linear parametrization (that is, $\tau(\mathbf{x}) = \tau^\top \mathbf{x}, \sigma(\mathbf{x}) = \sigma^\top \mathbf{x}, \alpha(\mathbf{s}) = \alpha^\top \mathbf{s}, \beta(\mathbf{s}) = \beta^\top \mathbf{s}$ for some real coefficients $\tau, \sigma, \alpha, \beta$ and their estimation is carried out via classical maximum likelihood) or non-parametric smooth estimation using splines (GAM, [51]), but any method can be used in practice. In case of a very small sample size, we can also assume a constant scale parameter $\sigma(\mathbf{x}) \equiv \sigma \in \mathbb{R}$ in order to reduce the dimension of the estimation.

The choice of q in the first step is a standard issue in extreme value theory [55–57]. For theoretical results, q should be growing with the sample size; that is, $q = q_n$ satisfying $\lim_{n\to\infty} q_n = 1$ and $\lim_{n\to\infty} n(1 - q_n) = \infty$. In practical terms, q should be set as high as possible while ensuring that a sufficient amount of data remains above the threshold to maintain good inferential properties. Classical choices include $q = 0.9$, $q = 0.95$, or $q = 0.99$, depending on the size of our dataset.

We utilize the basic bootstrap technique (sometimes also called Efron's percentile method, see Chapter 23 in [58]) to establish confidence intervals. This involves random sampling, with replacement, from our dataset to generate multiple bootstrap samples, each matching the size of our original dataset. For each bootstrap sample, we calculate the estimate of the statistic $\hat{\mu}^\star(t)$. Subsequently, we determine the α-percentiles of the re-sampled statistics to derive the confidence intervals. Details can be found in Appendix D.1.

Remark 2. *One must be cautious when interpreting confidence intervals during extrapolation. Generally, estimation of $\mu(t)$ is subject to two primary sources of bias: (1) bias stemming from model misspecification, and (2) bias arising from estimation variance; while the former bias can be mitigated within the body of the distribution by comparing different models and employing cross-validation, AIC or BIC criteria to select the most suitable model, this approach becomes less reliable in the extremal region. Eliminating this bias necessitates observation of data within the region of interest. The latter bias stemming from estimation uncertainty can be addressed by computing confidence intervals (in our case, via bootstrapping). It is important to note that our bootstrap confidence intervals only account for the latter bias, and consequently, the first type of bias presents a greater challenge during extrapolation since it is, in principle, unquantifiable without additional data.*

Theorem 2 (Idea: Precise statements can be found in Appendix D). *Assuming that the conditions outlined in either Theorem A1 or Theorem A2 are met, our procedure is consistent. Furthermore, under the assumptions detailed in Theorem A3, the bootstrap confidence intervals are asymptotically consistent at a correct level.*

In our procedure, we adopt a practice common in extreme value theory, where we concentrate solely on the extreme observations (set S) while discarding all non-extreme values. This approach stems from the rationale that extreme observations provide the most valuable insights into out-of-support behavior. Utilizing data within the body of the distribution may introduce bias, as these values may exhibit different behavioral patterns. Mathematically, this rationale can be expressed through an examination of the precision of the GPD approximation. This approximation shows high precision exclusively in extreme values while displaying bias and low precision for non-extreme values.

5. Illustration and Experiments

In this section, we assess the performance of our methodology using both a simple illustrative example and experimental data. A comprehensive simulation study is provided in Appendix B.

The quantity of interest in the application presented in Section 6 is the difference $\mu(t_1) - \mu(t_2)$ for $t_1, t_2 < \tau_R$. Hence, in the simulations, we focus on estimating

$$\omega_{\mathbf{x}} := \lim_{t \to \infty} [\mu_{\mathbf{x}}(t+1) - \mu_{\mathbf{x}}(t)] \quad \text{or} \quad \omega := \lim_{t \to \infty} [\mu(t+1) - \mu(t)],$$

assuming $\tau_R = \infty$ and that the corresponding limits exist. Note that under the linear model (5), the limit exists and corresponds to the parameter $\omega_{\mathbf{x}} = \beta[\theta(\mathbf{x})]$. Here, $\omega_{\mathbf{x}}$ can be regarded as the tail counterpart of a coefficient $\beta_{\mathbf{x}}$ in a linear model $Y = \alpha_{\mathbf{x}} + \beta_{\mathbf{x}} T + \varepsilon$, where $\alpha_{\mathbf{x}}$ and $\beta_{\mathbf{x}}$ are real coefficients, possibly dependent on $\mathbf{X}$.

5.1. Simple Example

The subsequent illustrative example outlines our methodology and the main ideas. Consider a single confounder $X = X_1 \sim \text{Bernoulli}(0.75)$ (where $X_1 = 1$ denotes men and $X_1 = 0$ denotes women, for instance). Define $T = X_1 + \varepsilon_T$, where $\varepsilon_T \sim \mathcal{N}(0, 1)$ (indicating that T generally tends to be larger for men than for women). Let

$$Y = \begin{cases} T + \varepsilon, & \text{when } X_1 = 1, \ T > 1, \\ 2T + \varepsilon, & \text{when } X_1 = 0, \ T > 1, \\ 2 - T + \varepsilon, & \text{when } T \leq 1, \end{cases}$$

where $\varepsilon \sim \mathcal{N}(0, 1)$.

Simple computation gives us $\mu(t) = 0.75t + (1 - 0.75)2t = 1.25t$ for any $t > 1$, while $\mu(t) = -t + 2$ for $t \leq 1$. Consequently, our primary interest lies in estimating the slope

$$\omega = \mu(t+1) - \mu(t) = 1.25 \quad \text{for } t > 1.$$

We generate data as specified with a sample size of $n = 500$. Setting the threshold at $q = 0.9$, we employ the methodology outlined in Section 4 to estimate ω. This process is repeated 100 times, yielding a mean and 0.95 quantile of

$$\hat{\omega} = 1.26 \pm 0.39.$$

Additionally, we employ the bootstrap technique to calculate confidence intervals, and we obtain a confidence interval of the form $\omega \in (0.72, 1.87)$ on average. We see from other simulations that these confidence intervals are slightly conservative for $n \le 1000$ but work well for larger sample sizes.

In Figure 2, we present one generated dataset with a sample size of $n = 500$, showcasing various methodologies from the existing literature and their extrapolation efficacy. Notably, classical techniques often exhibit a tendency to underestimate $\mu(t)$ for t large, primarily due to the fact that only the 'men' category ($X = 1$) received $T > 2$.
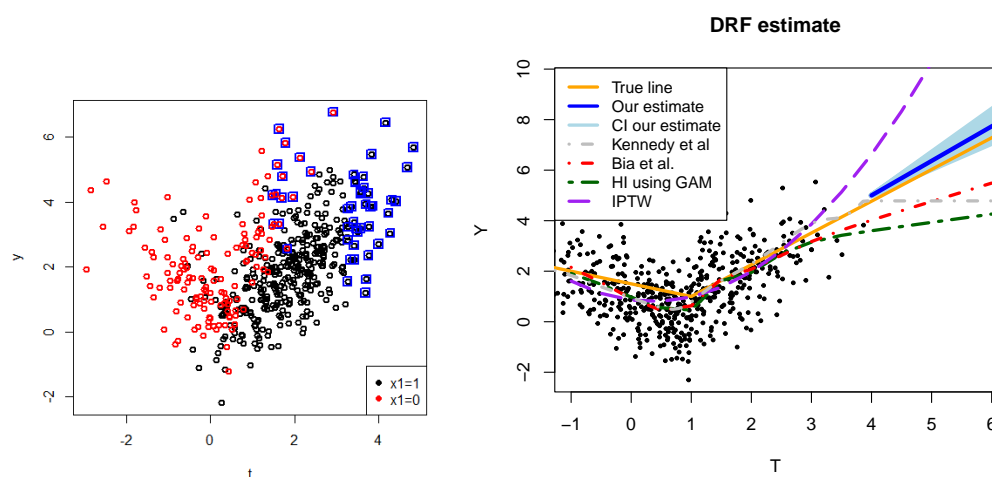
Figure 2. Left: Dataset generated based on the simulations outlined in Section 5.1 with $n = 500$. Points falling within the set S are identified by a blue square. **Right**: Estimation of $\mu(t)$ using various methods: orange represents the true $\mu(t)$, blue depicts our estimate employing the method from Section 4 with 95% confidence intervals, grey illustrates the doubly robust estimation method introduced by [11,59], red showcases the additive spline estimator described in [60], dark green demonstrates the approach proposed by [36] utilizing a GAM outcome model (further details in [61]), and purple describes the inverse probability of treatment weighting estimator [62].

We conclude with an important remark regarding the sample size: a substantial amount of valuable information is lost when we discard 90% of the data by focusing solely on the data in the set S (data above the threshold $\tau(\mathbf{x})$). This is the primary reason behind the considerably large confidence intervals and the heightened variability in our estimates. We encounter the inevitable bias-variance trade-off; the inclusion of more data introduces a potential bias, given that the behavior of $\mu(t)$ differs in the body and in the tail.

5.2. Simulations

We provide a comprehensive discussion of all simulations in detail in Appendix B. In our study, we devised several simulation setups to model diverse scenarios and explore them thoroughly. Specifically, we focused on the following five key scenarios:

1. Investigating how our method scales with respect to the dimension of the confounders $d = dim(\mathbf{X})$.
2. Comparing our method with classical methods from the literature.

3. Expanding upon the simple example introduced in Section 5.1, wherein we evaluated performance across various dependence structures (employing different copulas), sample sizes, and a spectrum of causal effects.
4. Examining the presence of a hidden confounder affecting both T and Y.
5. Focusing on variations in the function $\mu(t)$.

In this section, we present two key findings from our simulation study. Table 1 illustrates how our methodology scales with varying dimensions of the confounders $d = dim(X)$. Additionally, Table 2 depicts the comparison of our method with other classical methods from the literature, where we estimated $\mu(\tilde{t})$ for $\tilde{t} = \max_{i=1,\dots,n} t_i + 10$.

Table 1. Estimates of $\omega = -1$ with varying dimensions of the confounders $d = dim(X)$ and with different distributions of the noise of T. The sample size is $n = 5000$. The full simulations setup can be found in Appendix B.1.

True $\omega = -1$	Gaussian ε_T	Exponential ε_T	Pareto ε_T
$d = 5$	$\hat{\omega} = -1.0 \pm 0.03$	$\hat{\omega} = -1.0 \pm 0.01$	$\hat{\omega} = -1.0 \pm 0.001$
$d = 25$	$\hat{\omega} = -0.96 \pm 0.08$	$\hat{\omega} = -0.99 \pm 0.01$	$\hat{\omega} = -1.0 \pm 0.001$
$d = 50$	$\hat{\omega} = -0.75 \pm 0.28$	$\hat{\omega} = -0.97 \pm 0.09$	$\hat{\omega} = -0.99 \pm 0.01$
$d = 200$	$\hat{\omega} = -0.44 \pm 0.37$	$\hat{\omega} = -0.53 \pm 0.42$	$\hat{\omega} = -0.91 \pm 0.68$

Table 2. Comparing the extrapolation performance of various models using the average Absolute Relative Error (ARE) across 100 simulations with varying dimension of the confounders $d = dim(\mathbf{X})$. The interpretation of the values is as follows: if the true value of $\mu(\tilde{t})$ is 1, an ARE of 0.17 indicates an approximate typical error of $|\hat{\mu}(\tilde{t}) - \mu(\tilde{t})| \approx 0.17$. Bold values correspond to the algorithm with the best performance.

	Our Method	Bia et al. [60]	Kennedy et al. [11]	HI with GAM [36]	IPTW [62]
$d = 2$	**0.18**	0.68	0.64	0.42	3.89
$d = 10$	**0.48**	0.81	0.65	0.67	5.69
$d = 30$	**0.79**	0.92	could not handle	0.92	4.70

Table 1 suggests that the results are reasonably accurate as long as $d \leq 25$. As discussed in the Appendix B, the reason for the bias observed in larger dimensions d is that Assumption 1 and Lemma 2 are only asymptotic results, and with higher dimensions d, we require more data for the asymptotic theory for $T \mid \mathbf{X}$ to take effect. It is well known that the convergence rate of the maxima of the Gaussian random sample to an extreme value distribution is very slow, whereas it is faster with Exponential or Pareto distribution [63].

Analysis of Table 2 reveals that our method achieves the smallest extrapolation error. Hirano and Imbens [36] method utilizing a GAM outcome model showed surprisingly reasonable performance, while the IPTW method [62] exhibits poor performance due to the quadratic nature of the extrapolating curve. Please note that our method assumes linearity in the tail. If this assumption is not met, our method may produce inferior results. In such instances, alternative regression techniques can be utilized in step 2 of our algorithm instead of linear regression. For example, neural networks, as demonstrated in [34], can be employed to address nonlinear behavior and enhance performance.

6. Application: River Discharge Dataset

Understanding the causal relationship between extreme precipitation and river discharge is crucial for effective water resource management. In this study, we examine how extreme precipitation events impact river discharge. By utilizing a comprehensive dataset spanning various hydro-logical conditions, our research seeks to provide insights into the critical nexus between extreme precipitation dynamics and extreme

river discharge events. The data were collected by the Swiss Federal Office for the Environment (https://www.hydrodaten.admin.ch/ (accessed on 15 October 2023)) but were provided by the authors of [23,64], with some useful preliminary insights. We used precipitation data and other relevant measured variables from meteorological stations provided by Swiss Federal Office of Meteorology and Climatology, MeteoSwiss (https://gate.meteoswiss.ch/idaweb/login.do (accessed on 15 October 2023)).

We exclusively examine the discharge levels of the River Reuss, situated near Zurich in Switzerland (Figure 3). We selected this river due to the availability of excellent measurements of its discharge levels, complemented by well-documented weather conditions from nearby meteorological stations and diverse landscape. Our measurements include average daily discharges between January 1930 and December 2014 and daily precipitation in the nearby meteo-stations. 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.
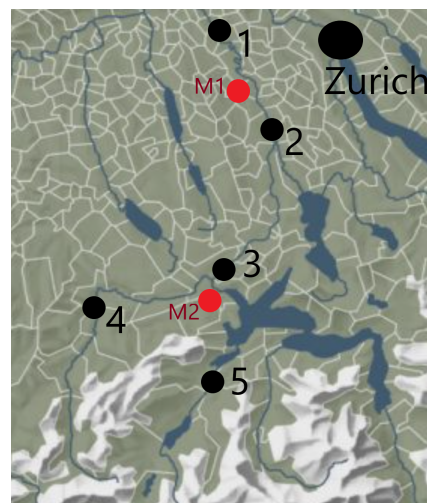
Figure 3. Map of meteo-stations (red) and five river stations (black). Note that the river flow is from south to north (with springs in the mountains).

We center our attention on addressing two distinct research questions: one characterized by a straightforward scenario where the ground truth is known, and another presenting a more intriguing challenge.

6.1. Known Ground Truth

We demonstrate our methodology using a straightforward example where the ground truth is known. Consider a pair of river stations, such as stations 2 and 1. Let T represent the water discharge at station 2, Y represent the water discharge at station 1, and $\mathbf{X}$ denote measurements taken at a nearby meteorological station (including precipitation, humidity, etc.; the full list of confounders can be found in Appendix C). Our objective is to investigate the impact of extreme discharge levels at station 2 on the water discharge observed at station 1. In mathematical terms, we seek to ascertain $\mu(t)$ or $\mu_{\mathbf{x}}(t)$ for large values of t. In this context, the ground truth is the following:

$$\mu(t_1 + t_2) - \mu(t_2) = \mu_{\mathbf{x}}(t_1 + t_2) - \mu_{\mathbf{x}}(t_2) = t_1 - t_2, \qquad \text{for all } t_1, t_2 \geq 0 \text{ and all } \mathbf{x} \in \mathcal{X}.$$

This can also be explained in words as follows: if we pour t_1 liters of water into the river at station 2 (in causal terminology, we interpret this as an intervention $do(T = T + t_1)$), we expect the water discharge at station 1 (Y) to increase by exactly t_1. Hence, $\omega = \omega_{\mathbf{x}} = 1$. As we will see below, our methodology consistently yields this expected outcome.

We follow the methodology introduced in Section 4 with $q = 0.95$. Detailed steps, diagnostics and preliminary data analysis (just for a pair $2 \to 1$) can be found in Appendix C.

The resulting estimates can be found in Table 3. The results are very similar with different choices of q (changing $\hat{\omega}$ by not more than by 0.1). We observe that our results align very well with the ground truth ($\omega = 1$). However, there is a slight bias evident in the relationships between the pairs $5 \to 3$ and $4 \to 3$: this can be attributed to distinct geographical features. Notably, Lake Vierwaldstättersee lies between the pair 5 and 3, which diminishes the influence of 5 on 3. Additionally, a 3238 m Titlis mountain is situated between pair 4 and 3, amplifying the effect of 4 on 3 due to the melting glacier ice, acting as an unmeasured confounding factor. Our methodology relies on Assumptions 1–4, along with some continuity assumptions and the SUTVA assumption discussed in Section 2. Assumptions 1 and 3 are minor and are used frequently when dealing with these types of data. Assumption 2 is a common and challenging aspect of every causal inference methodology; while our assumption is weaker than the classical unconfoundness assumption (requiring no hidden confounder in the tail), complete rejection of the possibility of its violation is unattainable. However, we believe that the meteo-station between a pair of stations can capture the most significant confounders (with the exception of when the lake or mountains are present in between the river stations). Finally, Assumption 4 is a strong assumption that allows us to extrapolate observed values into the extremal region. However, this assumption (or at least some similar model assumptions) are necessary. In this case, the linear assumption is valid, since the underlying ground truth is known.

Table 3. Estimates $\hat{\omega}$ between each pairs of the stations.

Truth: $\omega = 1$	Stations $2 \to 1$	Stations $3 \to 2$	Stations $4 \to 3$	Stations $5 \to 3$
$\hat{\omega}$	1.03 ± 0.05	1.17 ± 0.24	1.21 ± 0.19	0.78 ± 0.41

6.2. Effect of Precipitation on River Discharge

We employ our methodology to address a more complex inquiry where the ground truth is not known. Let us consider water discharge at station 3 (Y), and let T denote the precipitation measured in meteo-station M2. Our focus lies in understanding the impact of extreme precipitation events (T) on the water discharge (Y). As mentioned in the introduction, on 6 June 2002, we recorded a historical maximum precipitation level of $T_{\max} = 111 \frac{\text{mm}}{\text{m}^2}$, coinciding with the scenario when the river nearly breached its banks. Our inquiry centers on the question: how would the river discharge Y alter if T were to reach $120 \frac{\text{mm}}{\text{m}^2}$? In mathematical terminology, we are interested in estimating $\mu(120) - \mu(111)$ or possibly $\mu_{\mathbf{x}^\star}(120) - \mu_{\mathbf{x}^\star}(111)$, where $\mathbf{x}^\star$ are other covariates corresponding to that event. Addressing this question is challenging as we lack data within this extreme regime, necessitating reliance on extrapolation. This task is especially challenging, since we anticipate that the effect of precipitation on river discharge may vary between the body of the distribution and its tail, since the ground absorbs a significant portion of the rainfall during a light rain.

We follow the methodology introduced in Section 4. A straightforward approach would be to define T as precipitation and Y the water discharge on the same day, while choosing appropriate confounders $\mathbf{X}$ from some measurements at M2. Then, we can use the classical approach for estimating $\mu(t)$ in the body and our approach to estimate it in the tail. However, some problematic issues arise in this application:

- **Time issue:** $T_{monday} \to Y_{monday}$ but also $T_{monday} \to Y_{tuesday}$ since it takes time for the rain water to reach the river and rain tends to be more frequent around midnight. In fact, correlation (and extreme correlation coefficient as well, see Figure A11) is much higher for a pair ($T_{monday}, Y_{tuesday}$) than for (T_{monday}, Y_{monday}). The extreme storm on 6 June 2002 corresponded to extremely high river discharge on 7 June 2002 (where Y was about five times larger than on 6 June 2002). Hence, our interest lies in the effect $T_{monday} \to Y_{tuesday}$ (that is, we consider t_i as precipitation on day i while y_i is the discharge on day $i + 1$). Additionally, the presence of time introduces an autocorrelation issue. This can be handled by taking for example weekly maxima or discarding consecutive observations within a certain time frame to reduce the auto-

correlation effect. Alternatively, applying techniques like time series decomposition, differencing, or using autoregressive models can also mitigate the issue of auto-correlation in the data analysis process. We leave the data unchanged since the temporal dependence is primarily local, spanning only a few days, and does not introduce a substantial bias.

- **Variable selection issue:** choosing appropriate confounders **X** that act as confounders of Y and T. It is not clear which variables can be safely considered as confounders: if a variable X lie on a path $T \rightarrow X \rightarrow Y$, adjusting for X would lead to so-called path-canceling causal effect [65]. Here, we are interested in the so-called total causal effect, so we need to be cautious of which covariates to adjust for. However, not adjusting for a common cause leads to a bias. Moreover, there is often a feedback loop: $X_i \leftrightarrow Precipitation$ for X_i for example humidity or temperature. However, some of the variables can be safely considered as common causes: for example, the temperature on Sunday (the day before measuring precipitation). There is a huge amount of literature for such a variable selection, and we do not aim to comment on this research area—we only provide a full list of chosen confounders in Appendix C.

We estimate two values: $\hat{\omega}$ which is the tail quantity defined as the difference between $\mu(t+1) - \mu(t)$ for large t: in our case, how would Y change if it was raining by 1 $\frac{\text{mm}}{\text{m}^2}$ more on 6 June 2002? Next, we also estimate $\hat{\beta} = \mu(t+1) - \mu(t)$ corresponding to the body of the distribution (see Appendix C.2.2 for details on its computation). The resulting estimates can be found in Table 4 and visualization of the $\mu(t)$ can be found in Figure 4. We observe that the effect of T on Y is larger in the extreme region than in the body of the distribution by a factor of $\frac{3.04}{2.4} \approx 1.25$.

Table 4. Estimates $\hat{\beta}$ and $\hat{\omega}$ represent the estimation of the effect of T on Y in the body and in the tail, respectively. $\hat{\beta}$ is computed using standard regression, while $\hat{\omega}$ is the tail counterpart computed using steps introduced in Section 4. * Note that Station 5 is in different altitude and relatively far from meteo-station M2 with a lake in between; hence, there is a bias due to data collection problems.

Truth Unknown	Station 1	Station 2	Station 3	Station 4	Station 5
$\hat{\beta}$	2.4 ± 0.1	2.28 ± 0.1	1.44 ± 0.02	0.89 ± 0.02	0.38 ± 0.01
$\hat{\omega}$	3.04 ± 0.95	2.61 ± 0.67	1.62 ± 0.35	0.99 ± 0.32	0.36 ± 0.13 *
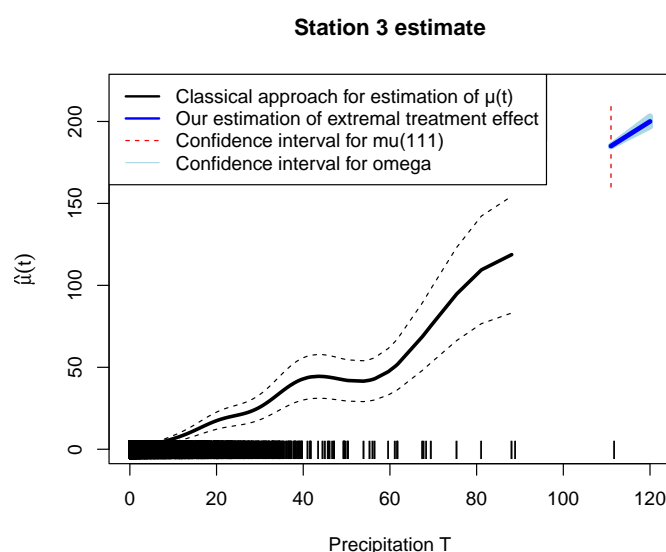
Figure 4. The estimation of $\mu(t)$ using the doubly robust estimator introduced in [11,59] cut at the second largest observation, its 95% confidence intervals, together with the estimation of $\mu(111), \omega$ and their 95% confidence intervals.

As for the answer to our question 'how would the river discharge Y alter in station 3 if T were to reach 120 $\frac{mm}{m^2}$ on 6 June 2002', our results suggest that the water discharge would be larger by about $9 \times 1.62 = 14.5 \frac{m^3}{s}$ (note that median of Y is 11.2, and the 95% quantile of Y is 51.2). Would this result in the river overflowing its banks? We cannot definitively say, as we lack the necessary data regarding the volume and contours of the river banks. Moreover, Y represents the daily average of the water discharge; while in order to answer this question, the daily maximum is a better suited variable for answering the question. Nonetheless, this advances us towards a more accurate understanding of the effects and impacts of extreme precipitation events and potentially enhancing statistical inference for hydroelectric power stations located along this river.

7. Conclusions and Future Work

Analyzing the impact of extreme levels of a treatment variable (exposure) is essential for comprehending its effects on diverse systems and populations. In this paper, we introduced a novel framework aimed at estimating the causal effect of extreme treatment values. Leveraging insights from extreme value theory, we enhanced the estimation of the extreme treatment effect. Our framework can handle a substantial number of confounders. Nonetheless, our methodology relies on extrapolation, presenting inherent challenges even in the absence of confounding variables, where the bias stemming from a model misspecification is impossible to quantify. Our framework holds promise for initial assessments of the impact of extreme environmental events, such as the effects of severe storms or droughts on economic damages. Future work may explore the application of our extreme value theory approach to address time-varying effects, a prevalent issue in environmental research.

Data Availability Statement: The code is available in the online repository (https://github.com/jurobodik/Extreme_treatment_effect.git (accessed on 25 March 2024)) or on request from the author, alongside with the data regarding Appendix A. While the data related to the application discussed in Section 6 are not publicly available, they can be accessed through https://www.hydrodaten.admin.ch/ (accessed on 15 October 2023) and https://gate.meteoswiss.ch/idaweb/login.do (accessed on 15 October 2023) after registration or by requesting the used data from the authors of [23].

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Application 2—Concrete Compressive Strength

Appendix A.1. Main Analysis

In this section, we delve into a dataset [66] focused on concrete compressive strength.

Concrete serves as a fundamental material in civil engineering, and understanding its compressive strength (denoted as Y and measured in MPa) is paramount for ensuring structural integrity [67]. Concrete comprises ingredients such as cement (X_1), fly ash (X_2), water (X_3), superplasticizer (X_4), and blast furnace slag (T) (among some other additions). The units of T, X_1, X_2, X_3, X_4 are in kilograms in a m^3 mixture. The concrete compressive strength (Y) exhibits a highly nonlinear relationship with these ingredients and the elapsed time. Our focus is on exploring the effect of blast furnace slag (T) on compressive strength (Y). It is well-established that increasing the quantity of T can enhance Y, yet an excessive amount of T may lead to a decrease in Y.

In our dataset X_1, X_2, X_3, X_4 may affect the decision of how much T was used (engineers often decide about the quantity of T based on the looks of the mixture of other ingredients). Our dataset contains $n = 1030$ instances of observational data $\{\mathbf{x}_i, t_i, y_i\}_{i=1}^n$ where $\mathbf{x}_i = (x_{1,i}, \ldots, x_{4,i})^\top$. The range $(\min_i y_i, \max_i y_i) = (2.3, 82.5)$ and $(\min_i t_i, \max_i t_i) = (0, 359)$.

Suppose we fit a linear model $\mathbb{E}Y = \beta_0 + \beta_T T + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$; then, a least square estimation of the coefficient β_T leads to $\hat{\beta}_T = 0.08 \pm 0.006$. This can be (wrongly) interpreted as 'adding one additional kg of T in m^3 mixture increases Y by $0.08 MPa$'. We expect different behavior for small and large values of T, and we expect strong (nonlinear) interactions between the covariates, and more importantly, this result is derived from the body of the distribution, while we are interested in values of T above the observed ones.

Our objective is to quantify the effect of an extreme amount of blast furnace slag T on Y. Specifically, we answer the following questions:

1. Given a concrete mixed with $T = 359$ and $\mathbf{X} = \mathbf{x}$ for some specific value of $\mathbf{x}$, if we intervene and change T to $T = 400$, what effect on concrete compressive strength can we expect? Using the potential outcome notation, the quantity of interest is $\mu_{\mathbf{x}}(400) - \mu_{\mathbf{x}}(359)$. Note that $max_{i=1,\ldots,n} t_i = 359$ (we do not observe the blast furnace slag larger than 359, and there is no observation in the interval $(220, 359)$), and hence, we have zero data in such an extreme region. We aim to answer this question for a choice $\mathbf{x} = \mathbf{x}^\star$ where $x_1^\star = 239, x_2^\star = 0, x_3^\star = 185, x_4^\star = 0$ (the observation corresponding to $T_i = 359$).

2. How would an extreme increase in T change Y for an 'average' concrete (on a population level, i.e., integrating over the covariates)? Using the potential outcome notation, the quantity of interest is $\mu(400) - \mu(359)$.

We follow the methodology introduced in Section 4 with $q = 0.9$. Detailed steps, diagnostics and preliminary data analysis can be found in Appendix A.2. The resulting estimates are as follows:

$$\hat{\mu}_{\mathbf{x}^\star}(400) - \hat{\mu}_{\mathbf{x}^\star}(359) = -4.1 \pm 3.0, \qquad \hat{\mu}(400) - \hat{\mu}(359) = -4.5 \pm 2.6,$$
$$\hat{\omega}_{\mathbf{x}^\star} = -0.1 \pm 0.07, \qquad\qquad \hat{\omega} = -0.11 \pm 0.06.$$

The results are similar with a different choices of q (see Table A1). In summary, the results suggest that for a mixture of concrete with covariates $\mathbf{X} = \mathbf{x}^\star$ and $T = 359$, intervening on T and changing it to $T = 400$ would decrease the concrete compressive strength by about 4.1 MPa. On the population level, increasing T from 359 to 400 would lead to a decrease in the concrete compressive strength by about 4.5 MPa. The 95% confidence intervals suggest that this estimate can be inaccurate by about 3 MPa; however, one must be cautious about the interpretation of the confidence intervals, since they are in general unreliable when extrapolating, see Remark 2.

Figure A1 graphically shows the estimation of $\hat{\mu}(t)$ in the body using the method introduced in [11,59], as well as our estimation of $\hat{\mu}(t)$ for extreme values.

In the Appendix A.4, we discuss the assumptions made regarding our results. In brevity, our methodology uses Assumptions 1–4 (among some continuity assumptions and SUTVA assumption discussed in Section 2). While we argue that Assumptions 1 and 3 are minor, we can not disregard the possibility of a hidden confounder (Assumption 2). The validity of Assumption 2 has to be further argued by an expert knowledge. Finally, the strongest assumption is Assumption 4, as its violation leads to the most significant bias. However, this assumption (or a similar assumption using different model) is necessary when extrapolating, and it is hypothetically testable by measuring values with $T \approx 400$. Appendix A.3 also discusses the differences for the range of choices of q.
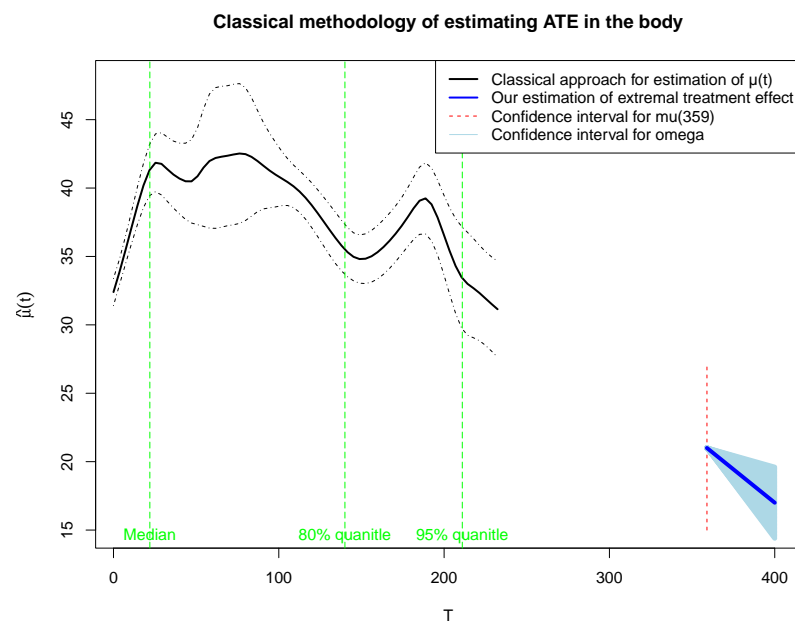
Figure A1. Black: The estimation of $\mu(t)$ using the doubly robust estimator introduced in [11,59] together with 95% confidence intervals. **Green:** Quantiles of T. **Blue:** Our estimation of $\mu(t)$ for values $t = 359, 400$ for $q = 0.9$, together with the 95% confidence intervals for the slope. **Red:** 95% confidence interval for $\mu(359)$.

Appendix A.2. Detailed Computations of the Estimates

Some data visualization can be found in Figures A2 and A4. In the following, we provide detailed descriptions of the steps undertaken in the application for the specific choice of $q = 0.9$. First, we estimate $\tau(\mathbf{x})$ using a classical quantile regression [68]. We observe that all covariates are highly significant, and the diagnostic plots do not show any significant problems (except the fact that for many observations, $T_i = 0$): we illustrate the estimation on Figure A5, where points above the 90% threshold (points in the set S) are marked.

In the next step, we routinely estimate $\theta(\mathbf{x})$ (using evgam function in evgam package [69] using the following code: evgam(list($T_e \sim s(X1_e) + s(X2_e) + s(X3_e) + s(X4_e), \sim 1$), $data = data.frame(T_e, X_e)$, family = "gpd") where T_e are datapoints in S (above the estimated 90% threshold)). More precisely, we assume fixed $\xi(\mathbf{x}) = \xi \in \mathbb{R}$ and only estimate $\sigma(\mathbf{x})$ as a smooth function of the covariates; Figure A3 shows the estimated values of $\sigma(\mathbf{x})$ on a log scale.

Finally, following the expression $\mathbb{E}[Y \mid T = t, \mathbf{X} = \mathbf{x}] = \alpha[\hat{\theta}(\mathbf{x})] + \beta[\hat{\theta}(\mathbf{x})]t$, we estimate α, β from the data points in S using gam function [51]. Under Assumptions 1–4 , we obtain $\hat{\mu}_{\mathbf{x}^\star}(400) = \hat{\alpha}[\hat{\theta}(\mathbf{x})] + \hat{\beta}[\hat{\theta}(\mathbf{x})]400$, and in effect, we return

$$\hat{\omega}_{\mathbf{x}^\star} = \hat{\mu}_{\mathbf{x}^\star}(400) - \hat{\mu}_{\mathbf{x}^\star}(359) = \hat{\beta}[\hat{\theta}(\mathbf{x}^\star)](400 - 359) = -4.1.$$

Regarding the second question (population level), we simply take the average $\hat{\mu}(t) := \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}[\hat{\theta}(\mathbf{x}_i)] + \hat{\beta}[\hat{\theta}(\mathbf{x}_i)]t$ and compute

$$\hat{\omega} = \hat{\mu}(400) - \hat{\mu}(359) = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}[\hat{\theta}(\mathbf{x}_i)](400 - 359) = -4.5.$$

Regarding the confidence intervals, we resample the data using the following code:

```
resampled_data = sample_n(data, size = length(y), replace = TRUE).
```

Then, we follow the same steps as above and estimate the coefficients (from the resampled dataset). We repeat this procedure 500 times. Finally, we take the 95% quantile out of all computed resampled coefficients. For example, $\hat{\omega}_{\mathbf{x}^{\star}} = -6.1 \pm 3.2$ represents the fact that the 95% quantile was $-6.1 + 3.2 = -2.9$, and hence, only 5% of values were larger than -2.9.
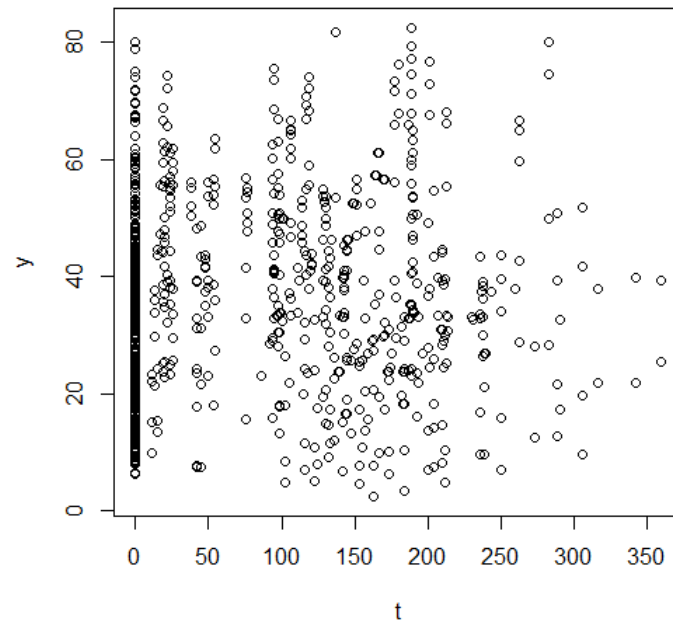
Figure A2. The figure illustrates the dependence among T and Y. Note that the correlation between T and Y is 0.14 ± 0.07.
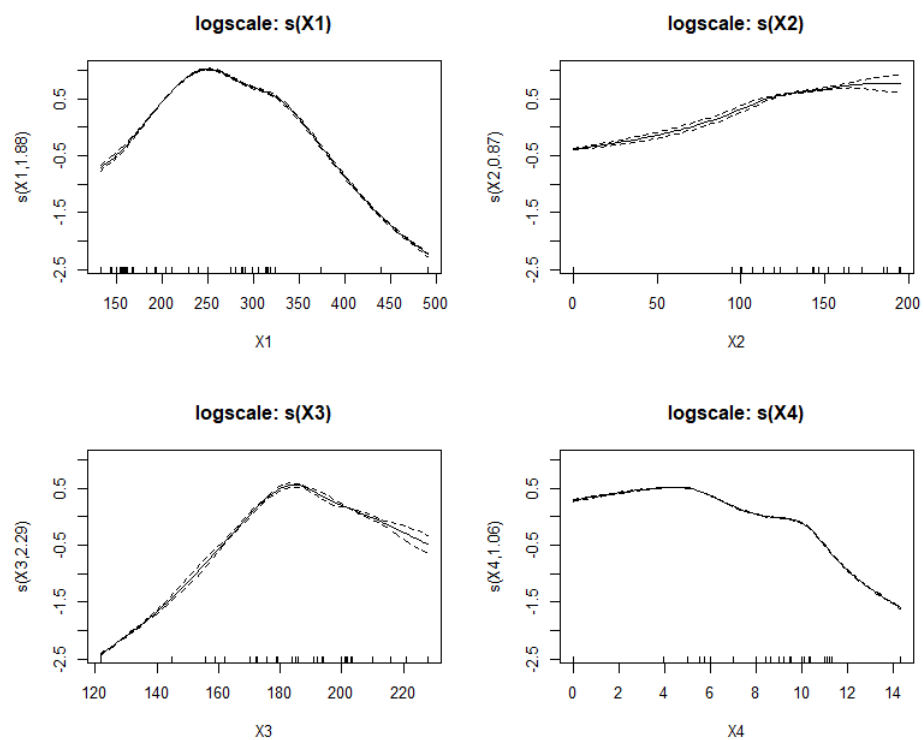
Figure A3. Estimation of scale.
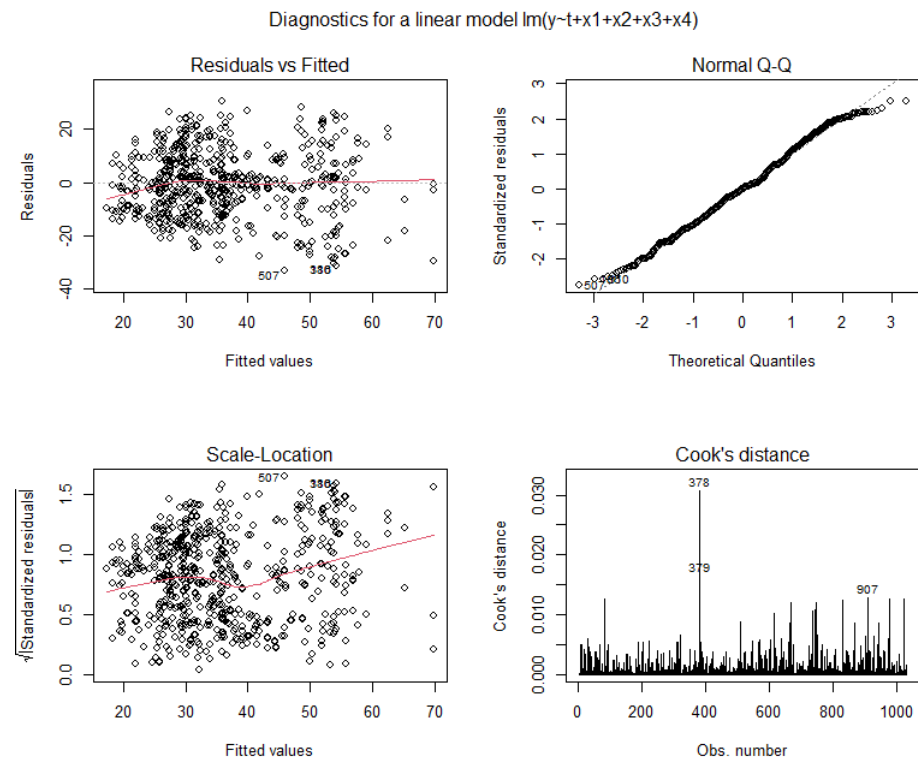
Figure A4. Diagnostics of a linear model fitted into the original data.
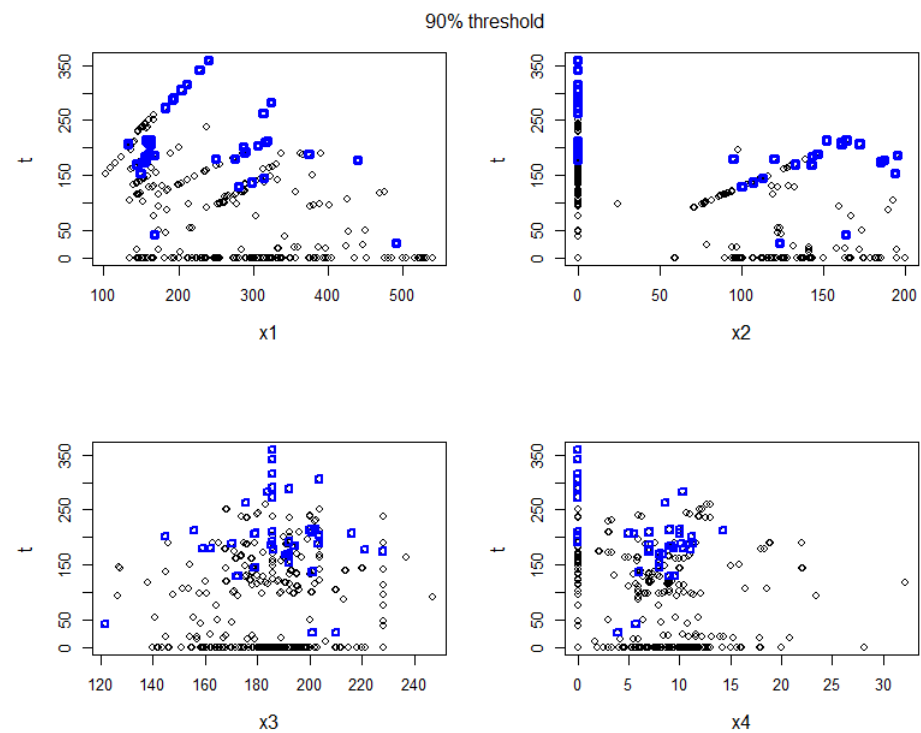
Figure A5. Visualization of the estimation of $\tau(\mathbf{x})$ estimated using classical quantile regression. Blue points characterize the observations above this threshold (points in the set S).

Appendix A.3. Discussion about the Results Regarding Different Threshold q

Table A1 shows the results for different choices of q. Even though they yield distinct estimates for $\hat{\omega}$, the confidence intervals overlap, and the values $\hat{\omega} \in (-6.4, -1.9)$ are encompassed by all of them. This suggests some stability in the choice of q.

Table A1. Estimates of $\omega_{\mathbf{x}^\star} := \mu_{\mathbf{x}^\star}(400) - \mu_{\mathbf{x}^\star}(359)$ for different thresholds q, together with the corresponding 95% confidence intervals.

$q = 0.85$	$q = 0.9$	$q = 0.95$
$\hat{\omega}_{\mathbf{x}^\star} = -6.1 \pm 3.2$	$\hat{\omega}_{\mathbf{x}^\star} = -4.1 \pm 3.0$	$\hat{\omega}_{\mathbf{x}^\star} = -2.8 \pm 2.8$
$\hat{\omega} = -5.3 \pm 4.0$	$\hat{\omega} = -4.5 \pm 2.6$	$\hat{\omega} = -3.3 \pm 3.1$

The selection of q reflects the bias–variance tradeoff; as we increase q, our inference relies on values closer and closer to $T = 400$ (datapoints with small and intermediate T_i can bias our estimation since in this region, increasing T_i can increase Y). However, increasing q also means disregarding more and more datapoints, and our estimate will have less power and larger variance.

The challenge of choosing q is a common problem in extreme value theory, and the rule of thumb is to select q as large as possible while maintaining an adequate number of datapoints above the $1 - q$ quantile to ensure reasonably good inference.

Appendix A.4. Discussion about the Assumptions

Our methodology relies on Assumptions 1–4, along with some continuity assumptions and the SUTVA assumption discussed in Section 2. Below, we provide a detailed discussion of each assumption.

1. Assumptions 1 and 3 are considered minor. As mentioned in Section 2, Assumption 1 is satisfied for most common distributions, and similar model assumptions are imposed in almost all applications utilizing extreme value theory. Assumption 3 appears to be satisfied, as there is no specific range of values in the support of T that has zero probability of occurring.

2. Assumption 2 is a common and challenging aspect of every causal inference methodology. While our assumption is weaker than the classical unconfoundness assumption (requiring no hidden confounder in the tail), complete rejection of the possibility of its violation is unattainable. A potential hidden confounder could be the 'quality of ingredients'. If the quality is low, engineers might tend to use excessive amounts of T in the mixture, potentially leading to spurious dependence between large T and low Y. However, in this case, it seems plausible that this hidden dependence due to low ingredient quality does not introduce a substantial bias. Expert knowledge is required to ensure the validity of this assumption.

3. Assumption 4 is a strong assumption that allows us to extrapolate observed values into the extremal region. However, this assumption (or at least some similar model assumptions) is necessary; estimating $\mu(400)$ from observed values is not feasible otherwise. In essence, Assumption 4 asserts that the relationship between T and Y (given other confounders) is linear in the unobserved region below $T = 400$. Since there is no other reason to believe that this relationship has any particular form, a linear assumption seems to be the most suitable choice. Although this assumption is strong, it is hypothetically possible to test by measuring values with $T \approx 400$.

Appendix B. Simulations

In this section, we create various simulations setups to assess the performance of our methodology.

*   Appendix B.1 provides insight into how our method scales with the dimension of the confounders $dim(\mathbf{X})$.
*   Appendix B.2 compares our method with classical methods from the literature.

- Appendix B.3 extends the simple example presented in Section 5.1, evaluating performance across different dependence structures (various copulas), sample sizes, and a range of causal effects.
- Appendix B.4 addresses a scenario involving a hidden confounder affecting both T and Y.
- Appendix B.5 focuses on variations in the function $\mu(t)$ and assesses the extent to which our method can extrapolate $\mu(t)$ into the 'extreme' region.

In some of the simulations, we use the following function:

$$\mu_x(t) = \begin{cases} 5 - slope(x)(t - c) & \text{for } t \geq c, \\ 5 + slope(x)(t - c) & \text{for } t < c, \end{cases} \tag{A1}$$

where typically $slope(x) = |x|$ and $c \in \mathbb{R}$ is a hyper-parameter. Graphical visualization of the function μ_x for $x = 3$ can be found in Figure A6. In other words, μ_x grows with slope x until c and then declines with slope x.
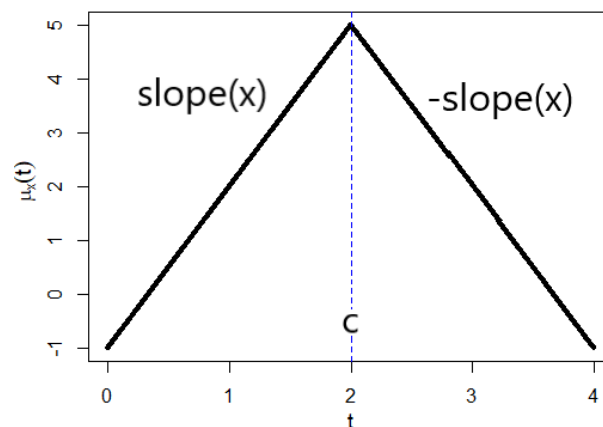
Figure A6. Function $\mu_x(t)$ with parameters $c = 2$ and $slope(x) = 3$.

Appendix B.1. Simulations with a High Dimensional $\mathbf{X}$

In this simulations we consider $\mathbf{X} = (X_1, \ldots, X_d)$ where the dimension d is potentially large. Consider the following data-generating process:

- Let $a_1, \ldots, a_d \overset{iid}{\sim} \mathcal{N}(1,1)$ and $b_1, \ldots, b_d \overset{iid}{\sim} \mathcal{N}(-1,1)$ be fixed numbers at the beginning of the simulations.
- Consider $\mathbf{X}$ being centered Gaussian vector with $cor(X_i, X_j) = 0.1$ for all $i \neq j$ and $var(X_i) = 1$.
- Let $T = \sum_{i=1}^d a_i X_i + \varepsilon_T$, where ε_T is distributed according to either $\mathcal{N}(0,10)$, $Exp(\frac{1}{10})$, or $Pareto(1,1)$.
- Let $Y = \mu_x(T) + \sum_{i=1}^d b_i X_i + \varepsilon_Y$, where $\mu_x(t)$ is defined in (A1) with hyper-parameters $c = slope(x) = 1$ and where $\varepsilon_Y \sim \mathcal{N}(0,1)$.

This data generating process leads to $\mu(t+1) - \mu(t) = -slope(x) = -1$ for $t \geq c$ and $\mu(t+1) - \mu(t) = +1$ for $t < c$. Consequently, our primary interest lies in estimating $\omega = \mu(t+1) - \mu(t) = -1$ for $t \geq c$.

Note that a simple linear regression $Y \sim T + X_1 + \cdots + X_d$ leads to a biased estimation of the effect of T, since the effect is different for large and for small values of T. However, simply discarding values where $T_i < 1$ leads to a selection bias.

We generate data as specified with a sample size of $n = 5000$. Setting the threshold at $\tau = 0.95$, we employ the methodology outlined in Section 4 to estimate ω. Specifically, we utilize linear parametrization of the parameters in the estimation procedure. This process is repeated 100 times. The mean of the estimates $\hat{\omega}$ together with 95% quantile for various values of d and distributions of the noise variables can be found in Table 1.

Table 1 illustrates that with a sample size of $n = 5000$, the results are reasonably accurate as long as $d \leq 50$. The reason for the bias observed in larger dimensions d is that Assumption 1 and (3) are only asymptotic results, and with higher dimensions d, we require more data for the asymptotic theory for $T \mid \mathbf{X}$ to take effect. It is well known that the convergence rate of the maxima of the Gaussian random sample to an extreme value distribution is very slow, whereas it is faster with the Exponential or Pareto distribution [63]. With a large dimension d, we also observe a more pronounced effect of the estimation error accumulated in the first step on the second step of the algorithm.

Appendix B.2. Comparison with Classical Methods

We evaluate our extrapolation method by comparing it with several state-of-the-art techniques from the existing literature. Specifically, we assess the performance of four methods: the doubly robust estimation method introduced by Kennedy et al. [11,59], the additive spline estimator proposed by Bia et al. [60], the approach suggested by Hirano and Imbens [36] employing a GAM outcome model (taken from [61]), and the inverse probability of treatment weighting (IPTW) estimator by VanderWal et al. [62].

Our analysis employs the same simulation setup described in Appendix B.1, utilizing exponentially distributed noise variables (other distributions yield similar results). After generating $(\mathbf{x}_i, t_i, y_i)_{i=1}^n$, we estimate $\mu(\tilde{t})$, where $\tilde{t} = \max_{i=1,\dots,n}(t_i) + 10$, using all aforementioned methods. Subsequently, we compute the absolute relative error (ARE) defined as

$$ARE = \left| \frac{\hat{\mu}(\tilde{t}) - \mu(\tilde{t})}{\mu(\tilde{t})} \right|.$$

This procedure is repeated 100 times, and the average of the obtained ARE values is presented in Table 2.

Analysis of Table 2 reveals that our method achieves the smallest extrapolation error. Conversely, the IPTW method [62] exhibits poor performance due to the quadratic nature of the extrapolating curve. Kennedy et al.'s method [11,59] typically produces a constant extrapolation curve, and while both Bia et al.'s [60] and the HI method [36] yield estimates that are reasonably close to the expected values, they also exhibit a notable degree of variability.

Please note that our method assumes linearity in the tail. If this assumption is not met, our method may produce inferior results. In such instances, alternative regression techniques can be utilized instead of linear regression. For example, neural networks, as demonstrated in [34], can be employed to address nonlinear behavior and enhance performance.

Appendix B.3. Dependence, Sample Size and the Causal Effect

In the following, we conduct simulations based on a model with covariates $\mathbf{X} = (X_1, X_2, X_3)$ that function as a common cause of both T and Y. The details of the simulation are as follows:

- $\mathbf{X}$ is generated with standard Gaussian margins and a Gumbel copula with a parameter α (where α represents the degree of dependence [70]; $\alpha = 1$ corresponds to independence; and $\alpha \to \infty$ corresponds to full dependence; see Figure A7 for an illustration with $\alpha = 2$).
- T is generated in such a way that the marginal distribution of T follows an exponential distribution with a scale parameter of 1, and the dependence structure between $\mathbf{X}$ and T follows a Gumbel copula with parameter α.
- The response variable Y is generated as follows:

$$Y = \begin{cases} \frac{1}{2}\omega T + f(\mathbf{X}) + \varepsilon, & \text{when } X_1 > 0, \ T > 1, \\ \frac{3}{2}\omega T + f(\mathbf{X}) + \varepsilon, & \text{when } X_1 \leq 0, \ T > 1 \\ -10T + 15 + f(\mathbf{X}) + \varepsilon, & \text{when } T \leq 1, \end{cases} \tag{A2}$$

where f is a randomly generated smooth function (to randomly generate a d-dimensional function, we use the concept of the Perlin noise generator [71]; for more details, refer to the supplementary package, and readers can conceptualize this as a function ranging from quadratic to linear), $\varepsilon \sim N(0,1)$, and ω is a hyper-parameter that we vary in our simulations. Figure A7 shows one realization of such a dataset.

Please note that $\mu(t) = \omega t + \mathbb{E}f(\mathbf{X})$ for any $t > 1$. As such, our primary focus lies in estimating the slope $\mu(t+1) - \mu(t) = \omega$. We generate data with varying parameters, including ω, α, and the sample size n. Employing the method outlined in Section 4, we estimate ω across a spectrum of data-generating processes. For $n = 1000$, we set the threshold at $\tau = 0.9$, and for $n > 1000$, we use $\tau = 0.95$. This process is repeated 100 times, and the mean and 0.95% quantile are presented in Table A2. The numbers in the brackets represent the mean of the estimated bootstrap confidence intervals. Ideally, these intervals should align with the 0.95% quantile.

The findings indicate that the methodology performs as anticipated in this simulation study: augmenting the sample size enhances the estimation, whereas elevating α (heightening the influence of the covariates) degrades the accuracy of the estimation. We observe that the bootstrap confidence intervals align relatively well with the actual 95% quantiles.
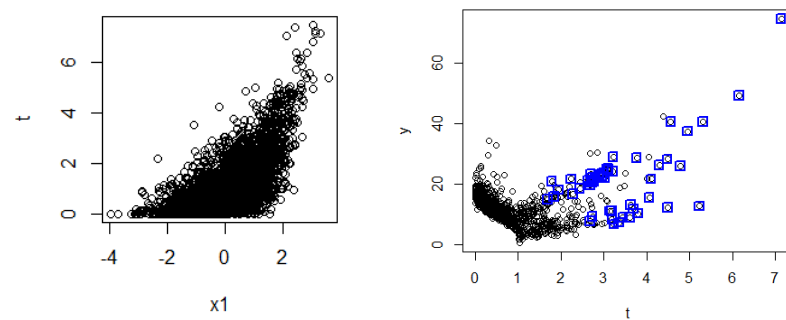
Figure A7. The figures illustrate the dependence among X_1, T, and Y, generated based on the simulations outlined in Appendix B.3 with a dependence parameter of $\alpha = 2$ and $\omega = 5$. Points falling within the set S are identified by a blue square.

Table A2. Resulting estimates of parameter $\omega = \mu(t+1) - \mu(t), t > 1$ from Appendix B.3. Parameter α represents the dependence between $\mathbf{X}$, T. The notation $\hat{\omega} = a \pm b \, (\pm c)$ represent the following: given 100 estimations of $\hat{\omega}$, a is the mean, b is the 95% quantile, and c is the (average) 95% quantile computed using the bootstrap technique.

		$\omega = 0$	$\omega = 1$	$\omega = 10$
	$n = 1000$	$\hat{\omega} = -0.05 \pm 0.33 \, (\pm 0.45)$	$\hat{\omega} = 0.94 \pm 0.60 \, (\pm 0.55)$	$\hat{\omega} = 9.86 \pm 2.99 \, (\pm 3.01)$
$\alpha = 1$	$n = 5000$	$\hat{\omega} = -0.01 \pm 0.28 \, (\pm 0.25)$	$\hat{\omega} = 0.98 \pm 0.29 \, (\pm 0.39)$	$\hat{\omega} = 10.09 \pm 1.91 \, (\pm 2.04)$
	$n = 10,000$	$\hat{\omega} = 0.00 \pm 0.18 \, (\pm 0.16)$	$\hat{\omega} = 0.99 \pm 0.21 \, (\pm 0.21)$	$\hat{\omega} = 9.95 \pm 1.62 \, (\pm 1.59)$
	$n = 1000$	$\hat{\omega} = 0.33 \pm 0.92 \, (\pm 0.96)$	$\hat{\omega} = 1.37 \pm 0.98 \, (\pm 1.13)$	$\hat{\omega} = 10.95 \pm 3.77 \, (\pm 3.38)$
$\alpha = 1.5$	$n = 5000$	$\hat{\omega} = 0.24 \pm 0.52 \, (\pm 0.68)$	$\hat{\omega} = 0.97 \pm 0.29 \, (\pm 0.25)$	$\hat{\omega} = 10.90 \pm 2.22 \, (\pm 2.43)$
	$n = 10,000$	$\hat{\omega} = 0.13 \pm 0.28 \, (\pm 0.49)$	$\hat{\omega} = 1.20 \pm 0.46 \, (\pm 0.55)$	$\hat{\omega} = 10.72 \pm 1.42 \, (\pm 1.76)$
	$n = 1000$	$\hat{\omega} = -0.17 \pm 1.19 \, (\pm 1.34)$	$\hat{\omega} = 0.99 \pm 1.01 \, (\pm 1.31)$	$\hat{\omega} = 11.14 \pm 3.32 \, (\pm 3.59)$
$\alpha = 2$	$n = 5000$	$\hat{\omega} = 0.03 \pm 0.66 \, (\pm 0.85)$	$\hat{\omega} = 1.05 \pm 1.01 \, (\pm 0.99)$	$\hat{\omega} = 11.15 \pm 2.91 \, (\pm 2.58)$
	$n = 10,000$	$\hat{\omega} = -0.09 \pm 0.50 \, (\pm 0.61)$	$\hat{\omega} = 0.96 \pm 0.59 \, (\pm 0.66)$	$\hat{\omega} = 10.70 \pm 2.02 \, (\pm 1.83)$

Appendix B.4. Simulations with a Hidden Confounder

Consider a similar simulations setup as in Section 5.1 but with a hidden confounder. Consider an observed confounder $X = X_1 \sim$ Bernoulli (0.75) and a hidden confounder

$H \sim N(1,1)$. Let $\delta \in \mathbb{R}$ and define $T = \delta H + X_1 + \varepsilon_T$, where $\varepsilon_T \sim \mathcal{N}(0,1)$. Note that δ represents the effect of a hidden confounder. Let

$$Y = \begin{cases} \delta H + \frac{2}{3}\omega T + \varepsilon, & \text{when } X_1 = 1, \ T > 1, \\ \delta H + \frac{6}{3}\omega T + \varepsilon, & \text{when } X_1 = 0, \ T > 1, \\ \delta H + 3 - 2T + \varepsilon, & \text{when } T \le 1, \end{cases} \tag{A3}$$

where $\varepsilon \sim \mathcal{N}(0,1)$. A simple computation leads to

$$\mu(t) = 0.75 \frac{2}{3}\omega t + (1 - 0.75)\frac{6}{3}\omega t = \omega t$$

for any $t > 1$, while $\mu(t) = -2t + 3$ for $t \le 1$. Consequently, our primary interest lies in estimating ω for $t > 1$.

We generate data as specified with a sample size of n. Setting the threshold at $\tau = 0.9$, we employ the methodology outlined in Section 4 to estimate ω. This process is repeated 100 times. The estimates $\hat{\omega}$ for a range of values of δ and ω and n can be found in Table A3.

The results in Table A3 suggest that a hidden confounder does not bias our estimate unless its strength is very large. Indeed, Remark 1 suggests that Assumption 2 is still valid since the hidden confounder enters the equality in an additive way.

Table A3. Resulting estimates of $\omega = \mu(t+1) - \mu(t), t > 1$ from Section B.4, together with 95% quantile. Parameter δ represent the strength of a hidden confounder.

		$\omega = 0$	$\omega = 5$	$\omega = 10$
	$n = 1000$	$\hat{\omega} = 0.05 \pm 0.22$	$\hat{\omega} = 5.03 \pm 0.31$	$\hat{\omega} = 10.02 \pm 0.4$
$\delta = 0$	$n = 5000$	$\hat{\omega} = -0.01 \pm 0.17$	$\hat{\omega} = 4.97 \pm 0.18$	$\hat{\omega} = 9.97 \pm 0.20$
	$n = 10,000$	$\hat{\omega} = 0.01 \pm 0.12$	$\hat{\omega} = 4.99 \pm 0.12$	$\hat{\omega} = 9.97 \pm 0.14$
	$n = 1000$	$\hat{\omega} = 0.55 \pm 0.27$	$\hat{\omega} = 5.60 \pm 0.32$	$\hat{\omega} = 10.48 \pm 0.33$
$\delta = 1$	$n = 5000$	$\hat{\omega} = 0.53 \pm 0.12$	$\hat{\omega} = 5.50 \pm 0.18$	$\hat{\omega} = 10.51 \pm 0.18$
	$n = 10,000$	$\hat{\omega} = 0.50 \pm 0.08$	$\hat{\omega} = 5.48 \pm 0.09$	$\hat{\omega} = 10.46 \pm 0.16$
	$n = 1000$	$\hat{\omega} = 0.96 \pm 0.04$	$\hat{\omega} = 5.98 \pm 0.22$	$\hat{\omega} = 10.91 \pm 0.21$
$\delta = 5$	$n = 5000$	$\hat{\omega} = 0.96 \pm 0.03$	$\hat{\omega} = 5.95 \pm 0.07$	$\hat{\omega} = 10.94 \pm 0.16$
	$n = 10,000$	$\hat{\omega} = 0.96 \pm 0.025$	$\hat{\omega} = 5.94 \pm 0.04$	$\hat{\omega} = 10.92 \pm 0.08$
	$n = 1000$	$\hat{\omega} = 0.98 \pm 0.06$	$\hat{\omega} = 5.96 \pm 0.15$	$\hat{\omega} = 10.94 \pm 0.30$
$\delta = 10$	$n = 5000$	$\hat{\omega} = 0.99 \pm 0.025$	$\hat{\omega} = 5.98 \pm 0.07$	$\hat{\omega} = 10.97 \pm 0.14$
	$n = 10,000$	$\hat{\omega} = 0.99 \pm 0.02$	$\hat{\omega} = 5.97 \pm 0.04$	$\hat{\omega} = 10.95 \pm 0.09$
	$n = 1000$	$\hat{\omega} = 0.99 \pm 0.01$	$\hat{\omega} = 5.99 \pm 0.14$	$\hat{\omega} = 10.97 \pm 0.28$
$\delta = 50$	$n = 5000$	$\hat{\omega} = 0.99 \pm 0.01$	$\hat{\omega} = 5.99 \pm 0.07$	$\hat{\omega} = 10.98 \pm 0.15$
	$n = 10,000$	$\hat{\omega} = 1.0000 \pm 0.004$	$\hat{\omega} = 5.97 \pm 0.04$	$\hat{\omega} = 10.95 \pm 0.08$

Appendix B.5. Simulations with Varying Extremal Region

In the following simulations, we explore variations in the function $\mu(t)$ and analyze the corresponding estimations $\hat{\mu}(t)$ for large values of t.

Consider the following data-generating process:

$$\begin{aligned} X &= \varepsilon_X, & \varepsilon_X &\sim N(0,1) \\ T &= X + \varepsilon_T, & \varepsilon_T &\sim t_\nu, \\ Y &= \mu_X(T) + \varepsilon_Y, & \varepsilon_Y &\sim N(0,1), \end{aligned}$$

Student's t distribution with ν degrees of freedom. Note that if $\nu \to \infty$, we obtain a Gaussian distribution, where $\mu_x(t)$ is defined in (A1). If c is too large, we only observe the region where μ_x grows, and hence, our estimation tends to be larger than the true value.

With varying c and ν, we estimate the parameter

$$\omega = \mu(t+1) - \mu(t) = -\mathbb{E}|X| = -0.798 \qquad \text{for } t \geq c.$$

If we fit a linear model $\mathbb{E}Y = \beta_0 + \beta_T T + \beta_X X$, the estimate $\hat{\beta}_T$ tends to be positive (depending on c and ν, for example, if $c = \nu = 5$, then $\hat{\beta}_T = 0.58 \pm 0.02$). Using our methodology, we estimate $\hat{\omega}$ as in the previous simulations. The resulting numbers are presented in Table A4. We observe that if c grows, our estimate becomes more biased as the data above the threshold still fall below $T < c$. Specifically, if $\nu = \infty$, only 0.2% of data points have $T > 5$, making the behavior of $\mu(t)$ above $t > c$ challenging to estimate. Note that the degrees of freedom ν correspond to the heavy-tailness of T; smaller ν values lead to more extreme values of T. Conversely, if $\nu = \infty$, T follows a Gaussian distribution. Heavier tails of T lead to better estimates.

Table A4. Estimates of $\hat{\omega}$ with varying c and ν. Note that true $\omega = -\mathbb{E}|X| \approx -0.79$.

| **True $\omega \approx -0.79$.**	**$c = 1$**	**$c = 2$**	**$c = 5$**	**$c = 10$**
$\nu = \infty$	$\hat{\omega} = -0.75 \pm 0.15$	$\hat{\omega} = -0.33 \pm 0.13$	$\hat{\omega} = 0.63 \pm 0.13$	$\hat{\omega} = 0.71 \pm 0.38$
$\nu = 5$	$\hat{\omega} = -0.78 \pm 0.18$	$\hat{\omega} = -0.72 \pm 0.15$	$\hat{\omega} = -0.13 \pm 0.15$	$\hat{\omega} = 0.6 \pm 0.33$
$\nu = 2$	$\hat{\omega} = -0.77 \pm 0.23$	$\hat{\omega} = -0.76 \pm 0.22$	$\hat{\omega} = -0.7 \pm 0.22$	$\hat{\omega} = -0.51 \pm 0.19$

Appendix C. River Data Application

Appendix C.1. Simple Illustration with Known Ground Truth

We used the following set of confounders:

- $X_1 =$Total precipitation (daily);
- $X_2 =$Total precipitation during the previous 7 days;
- $X_3 =$Daily maximum of air temperature 2 m above ground;
- $X_4 =$Daily maximum of relative air humidity 2 m above ground;
- $X_5 =$Daily mean of vapor pressure 2 m above ground;
- $X_6 =$Daily maximum of pressure reduced to sea level;
- $X_7 =$Daily total of reference evaporation from FAO.

For the pair $2 \to 1$, we considered measurements from meteo-station M1 (station code MURI, AG), and for the remaining pairs, we used measurements from M2 (a station with a code LUZ). However, variables X_5, X_6, X_7 were also measured in station M2 also for the pair $2 \to 1$, since some values were missing, and M2 has a much longer time period of measurements. All of these covariates can be safely considered as common causes of T and Y, and no feedback loop is present. For modeling of $\theta(\mathbf{X})$, we used linear parametrization, that is, $\tau(\mathbf{X}) = const + \sum_{i=1}^{7} \beta_{i,\tau} X_i$ and $\sigma(\mathbf{X}) = const + \sum_{i=1}^{7} \beta_{i,\sigma} X_i$, where the parameters $\beta_{i,\tau}$ were estimated using quantile regression (in R using `quantreg` package) and the parameters $\beta_{i,\sigma}$ were estimated using `evgam` package. We fixed $\xi(\mathbf{X})$ to be constant.

In the following, we focus on the pair $2 \to 1$; for other pairs of stations, the results were similar. In the modeling of $\theta(\mathbf{X})$, X_3 and X_5 were not significant at the 0.05 level (note that for an estimation of ω, we do not care which covariates were significant since the function $\theta(\mathbf{X})$ is more or less unchanged and adding non-significant covariates only slightly increases variance in the estimation). Using non-parameteric GAM estimation of the parameters did not change the final estimation much (from $\hat{\omega} = 1.03$ to $\hat{\omega} = 0.99$). Estimation of $\tau(\mathbf{X})$ is also plotted in Figure A8. Using the linear model in this case does not seem to be very wrong, see plot Figure A9, where except for the normality violation, the model seems to fit quite well (in the body of the distribution).
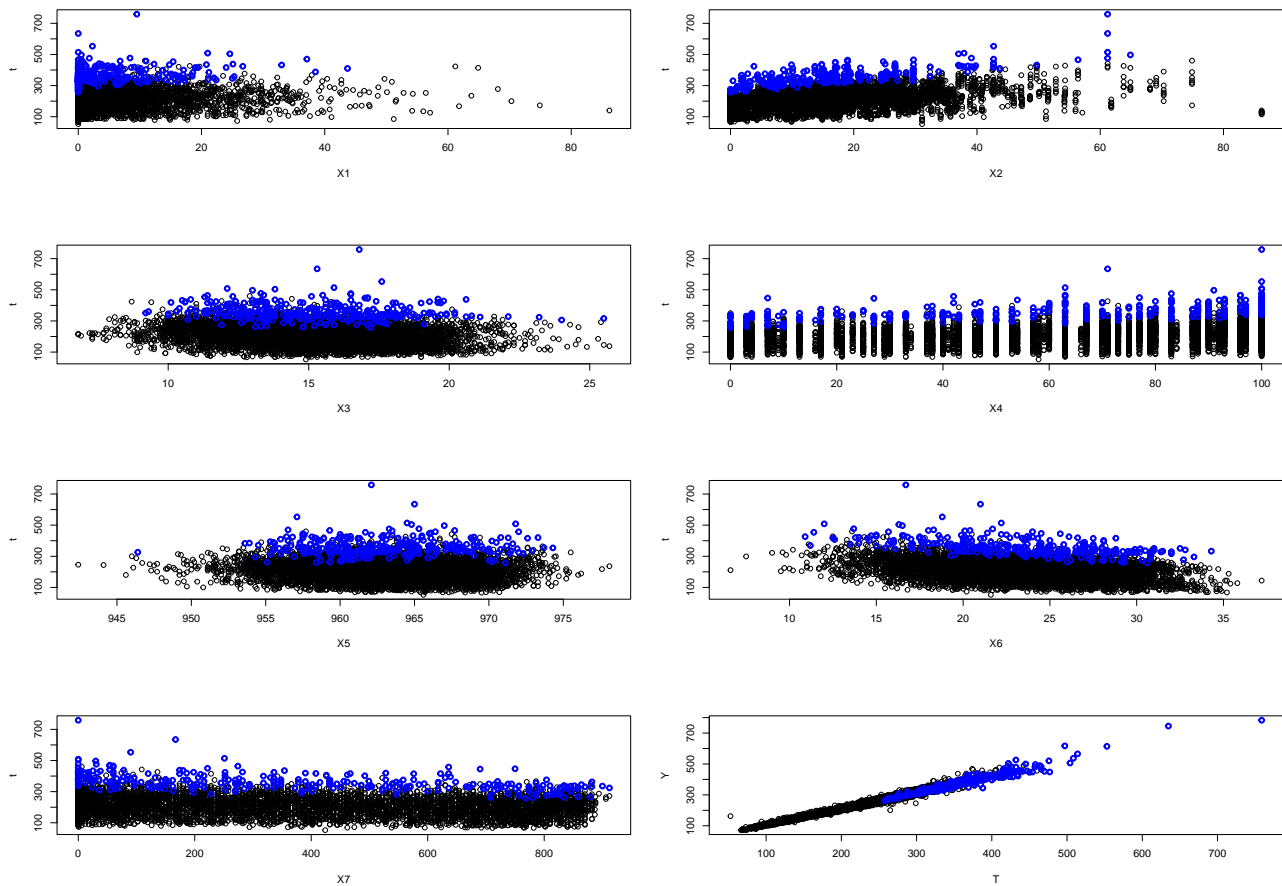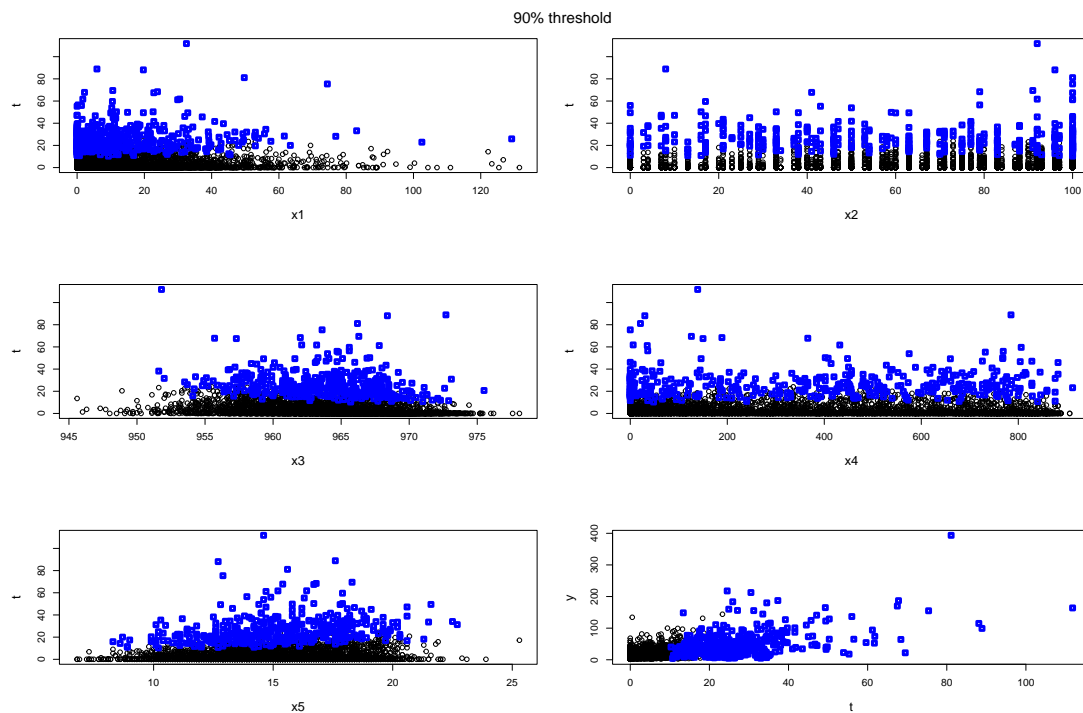
Figure A8. Visualization of the estimation of $\tau(\mathbf{x})$, estimated using classical quantile regression. Blue points characterize the observations above this threshold (points in the set *S*).
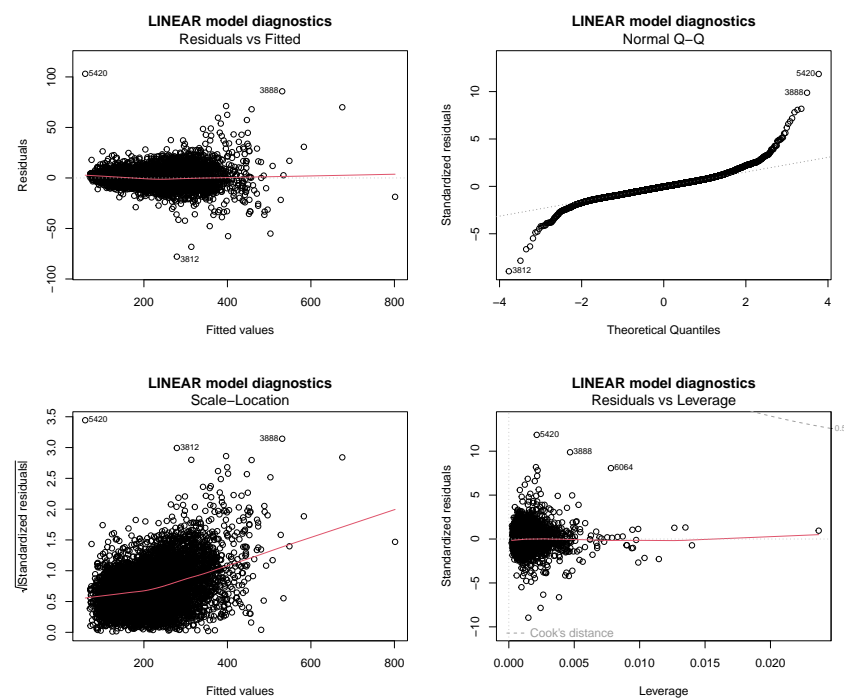
Figure A9. Diagnostics of a model $Y \sim X_1 + \cdots + X_7$.

Appendix C.2. Effect of Precipitation on River Discharge

Figure A10 visualizes the dataset.

Figure A10. Visualization of the estimation of $\tau(\mathbf{x})$ for Part2 in the application, estimated using classical quantile regression. Blue points characterize the observations above this threshold (points in the set *S*).

Appendix C.2.1. Choice of Variables

We used the following set of variables:

- $Y =$ River discharge on day $i + 2$;
- $T =$ Precipitation in the corresponding meteo-station on day $i + 1$;
- $X_1 =$ Total (sum) precipitation during the previous 7 days (days $i, i - 1, \ldots, i - 6$);
- $X_2 =$ Daily maximum of Air temperature 2 m above ground on day i;
- $X_3 =$ Daily maximum of Relative air humidity 2 m above ground on day i;
- $X_4 =$ Daily maximum of Pressure reduced to sea level on day i;
- $X_5 =$ Daily total of Reference evaporation from FAO on day i.

Here, i spans from 1.6.1930 up to 29.8.2014 (recall that we only considered the summer months). The choice of Y and T was addressed in the main text: since typically the autocorrelation peaking when T represents the day prior to Y (as illustrated in Figure A11, as well as its extreme counterpart extremogram [72]). The rationale behind choosing X_1 is straightforward—precipitation over the preceding days emerges as a significant confounding factor affecting both Y and T. Regarding additional variables, we opted for those deemed relevant and with reliable measurements across meteorological stations, all of which were recorded on the preceding day.

Is there a common cause between Y and T that remains unaccounted for? Variables X_2 through X_5, measured on day $i + 1$, could serve as potential common causes for both Y and T. For instance, a *sudden* temperature change might elevate the likelihood of intense rainfall, while alterations in river discharge could stem from specific soil characteristics affected by the temperature change. Nevertheless, we contend that the majority of these variables require more than a day to manifest their effects, which we believe are largely encapsulated by our chosen variables.
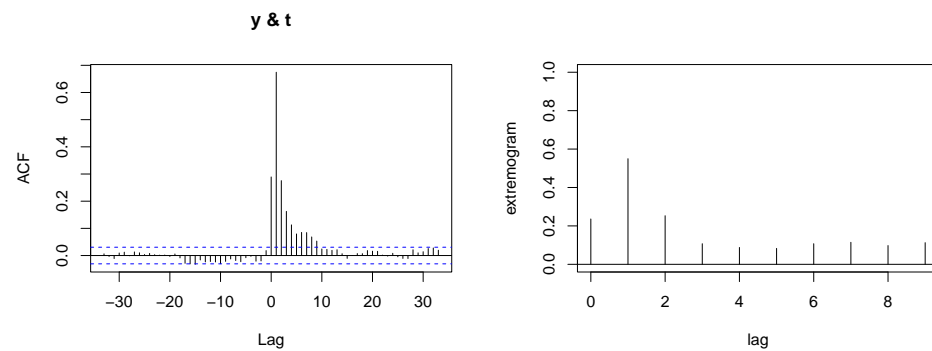
Figure A11. Cross-correlation and cross-extremogram of precipitation recorded at meteo-station M2 and water discharge at Station 3, both measured on the same day.

Appendix C.2.2. Computation of $\hat{\beta}$

In Table 4, we introduced a variable $\hat{\beta}$ that represents the effect of precipitation on the river discharge level in the body of T. This can be defined in several ways:

1. Using the method introduced in [11,59], we estimate $\hat{\mu}(t+1) - \hat{\mu}(t)$ for $t = \mathbb{E}(T)$.
2. Using a very straightforward approach where we model the data generating process of Y using a linear structural equation model $Y = c + \beta_T T + \beta_{\mathbf{X}} \mathbf{X} + \varepsilon$ and return the least square estimate of $\hat{\beta}_T$.

Coincidentally, both approaches return a very similar value of $\hat{\beta}$, and hence, it does not matter which approach we use (values in Table 4 are from using the second approach).

Appendix D. Consistency, Bootstrap and Its Asymptotics

In this section, we give a more detailed description of the bootstrap algorithm and a more precise statement of Theorem 2 together with its proof. Theorem A1 presents the consistency of $\hat{\mu}(t)$ for large t, while Theorem A2 shows the consistency of $\hat{\omega}_{\mathbf{x}^\star}$ under different assumptions. Note that using the notation from Sections 4 and 5,

$$\hat{\omega}_{\mathbf{x}^\star} = \lim_{t \to \infty} \hat{\mu}_{\mathbf{x}^\star}(t+1) - \hat{\mu}_{\mathbf{x}^\star}(t) = \lim_{t \to \infty} \hat{\alpha}[\hat{\theta}(\mathbf{x}^\star)] + \hat{\beta}[\hat{\theta}(\mathbf{x}^\star)](t+1) - \hat{\alpha}[\hat{\theta}(\mathbf{x}^\star)] - \hat{\beta}[\hat{\theta}(\mathbf{x}^\star)]t$$
$$= \hat{\beta}[\hat{\theta}(\mathbf{x}^\star)].$$
(A4)

Appendix D.1. Bootstrap

In what follows, we explain in detail the procedure for an estimator $\hat{\zeta}_\alpha$ satisfying

$$P(\omega_{\mathbf{x}^\star} \leq \hat{\zeta}_\alpha) \geq 1 - \alpha, \qquad \alpha \in (0, 1).$$

We only focus on the upper confidence interval, the lower and both-sided intervals can be carried out analogously. Our approach is standard and [58] provides a good overview.

Let P_n be the empirical distribution of the observations $\mathbf{Z}_i := (\mathbf{X}_i, T_i, Y_i), i = 1, \ldots, n$. We draw a random sample $(\mathbf{Z}_1^\star, \ldots \mathbf{Z}_n^\star) \overset{iid}{\sim} P_n$, and we compute the parameter $\hat{\omega}^\star$ from $(\mathbf{Z}_1^\star, \ldots \mathbf{Z}_n^\star)$ the same way as we compute $\hat{\omega}$ from $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. We define $\hat{\zeta}_\alpha$ as the upper α-quantile of $\hat{\omega}^\star$, that is, the smallest value $x = \hat{\zeta}_\alpha$ that satisfies

$$P\left(\hat{\omega}^\star - \hat{\omega} \leq x \mid P_n\right) \geq 1 - \alpha.$$

The notation $P(\cdot | P_n)$ indicates that the distribution of $\hat{\omega}^\star$ must be evaluated assuming that the observations are sampled according to P_n given the original observations. In particular, in the preceding display $\hat{\omega}$ is to be considered nonrandom.

It is almost never possible to calculate the bootstrap quantiles exactly [58]. In practice, these estimators are approximated by a simulation procedure. A large number of independent bootstrap samples $\mathbf{Z}_1^{\star}, \ldots, \mathbf{Z}_n^{\star}$ are generated according to the estimated distribution P_n. Each sample gives rise to a bootstrap value $\hat{\omega}^{\star}$. Finally, the bootstrap quantiles are estimated by the empirical quantiles of these bootstrap values. This simulation scheme always produces an additional (random) error in the coverage probability of the resulting confidence interval. In principle, this error can be made arbitrarily small by using a sufficiently large number of bootstrap samples. Therefore, the additional error is usually ignored in the theory of the bootstrap procedure. This section follows this custom and concerns the "exact" quantiles, without taking the simulation error into account.

Appendix D.2. Simplifying Assumptions

We simplify some steps in the inference process in order to simplify the proof of the consistency. In particular, we assume the following:

(A) **(Causality justification)** Consider Assumptions 1, 2 and 4 to be valid.
(B) **(Step 2 convergence)** $\mathbb{E}Y^2 < \infty$, $\mathbb{E}||\mathbf{X}||^2 < \infty$ and $(\mathbf{X}, T)$ satisfy Grenander conditions (this is a minor assumption assuring that the matrix of observations have a full rank with probability tending to one. See Table 4.2 in [73]).
(C) **(Step 1 convergence)** We assume that conditions R1, R2, and R3 from [74] are satisfied. That is, $E(\mathbf{X}\mathbf{X}^{\top})$ is positive semi-definite, $\mathbf{X}$ has a compact support with existing and finite quantile densities $\frac{\partial F_U^{-1}(\tau|x)}{\partial \tau}$, $\frac{\partial F_U^{-1}(\tau)}{\partial \tau}$ where $U = T - \tau_{lin}^{\top}\mathbf{X}$, $\tau_{lin} \in \mathbb{R}^d$.
(D) **(Linearity)** Assume that functions θ, α, β are linear, functions σ, ξ are constant and that we employ linear regression for the estimation of the parameters.

In particular, following the notation in Section 4 and using the notation $\tau(\mathbf{x}) = \tau_{lin}^{\top}\mathbf{x}$, our algorithm is as follows:

- Choose $q \in (0, 1)$,
- **(Step 1)** Estimate $\tau_{lin} \in \mathbb{R}^d$ by minimizing $\hat{\tau}_{lin} \in argmin_b \sum_{i=1}^n h_q(T_i - \mathbf{X}_i^{\top}b)$ where $h_q(x) = x(q 1_{x \geq 0} - (1-q)1_{q<0})$.
- **(Step 2)** We estimate $\alpha, \beta \in \mathbb{R}$ using least squares in a model

$$\mathbb{E}[Y \mid T = t, \tau(\mathbf{X}) = \hat{\tau}_{lin}^{\top}\mathbf{x}] = \alpha \hat{\tau}_{lin}^{\top}\mathbf{x} + \beta \hat{\tau}_{lin}^{\top}\mathbf{x}t, \tag{A5}$$

from the data-points in S (that is, we only consider $t > \hat{\tau}_{lin}^{\top}\mathbf{x}$). Using **R** language we run the following code: `fit = lm(Y ~ s +s:T_S, data = data.frame(s, T_S))`, where $s = \hat{\tau}_{lin}^{\top}\mathbf{X}_S$, $T_S = \{T_i : i \in S\}$ and $\mathbf{X}_S = \{\mathbf{X}_i : i \in S\}$.
- We output $\hat{\omega}_{\mathbf{x}^{\star}} = \hat{\beta}\hat{\tau}_{lin}^{\top}\mathbf{x}^{\star}$ (see (A4)).

Remark A1. *Assumption C implies consistency of $\hat{\tau}_{lin}$ (under the assumption that q is chosen as a function of the sample size n, denoted as $q = q_n$ satisfying $\lim_{n \to \infty} q_n = 1$ and $\lim_{n \to \infty} n(1 - q_n) = \infty$), see Theorem 5.1 in [74]. This assumption can be simplified by directly assuming consistency of $\hat{\tau}_{lin}$.*

Appendix D.3. Consistency

We present two consistency results. One concerns the consistency of $\hat{\mu}(t)$ under general non-linear assumptions but under a neglecting the GPD approximation error. The second result describes the consistency of $\hat{\omega}_{\mathbf{x}^{\star}}$ under linear assumptions presented in Appendix D.2.

Theorem A1 (Consistency). *Consider Assumptions 1, 2, and 4 to be valid.*

- *Assume that θ, α, and β are continuous functions, and suppose we employ consistent estimators for θ, α, and β. For instance, the Generalized Additive Model (GAM) estimator [51] has been shown to be consistent under specific smoothness conditions.*

- *Let $q \in (0,1)$ be chosen such that the distribution of $T \mid T > \tau_q(\boldsymbol{x}), \boldsymbol{X} = \boldsymbol{x}$ follows $GPD(\theta(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X} = supp(\boldsymbol{X})$ is assumed to be compact.*

  Under these conditions, our estimator is consistent in the sense that for all $t \in \mathcal{T}$

$$\hat{\mu}(t) \overset{P}{\to} \tilde{\mu}(t) \qquad \text{as } n \to \infty,$$

where $\tilde{\mu}$ is a function that satisfies $\tilde{\mu}(t) \sim \mu(t)$ as $t \to \tau_R$.

The second assumption outlined in Theorem A1 is introduced to address certain technical hurdles that arise when dealing with a quantile q which varies with the sample size n. Broadly speaking, when q is not fixed, the statistical framework becomes considerably more intricate, making the task of demonstrating the consistency of a quantile regression notably challenging. Please note that while the distribution of $T \mid T > \tau_q(\mathbf{x}), \boldsymbol{X} = \mathbf{x}$ converges to a Generalized Pareto Distribution ($GPD(\theta(\mathbf{x}))$) in the limit for large q, the exact validity is limited to special cases, such as when $T \mid \boldsymbol{X}$ follows a Pareto distribution. However, by selecting q to be sufficiently large, one can mitigate this issue, effectively reducing the disparity between the distributions of $T \mid T > \tau_q(\mathbf{x}), \boldsymbol{X} = \mathbf{x}$ and $GPD(\theta(\mathbf{x}))$ to insignificance. This theorem, therefore, provides valuable insight into the general consistency of the model, despite the idealized nature of the assumption.

Note that Theorem A1 can be reformulated for $\mu_{\boldsymbol{X}}$ analogously.

The subsequent theorem does not necessitate a fixed q; however, it presupposes the linearity in the models for T and Y.

Theorem A2 (Consistency). *Under Assumptions A, B, C, and D, where q is chosen as a function of the sample size n, denoted as $q = q_n$ satisfying $\lim_{n \to \infty} q_n = 1$ and $\lim_{n \to \infty} n(1 - q_n) = \infty$, our estimator $\hat{\omega}_{\boldsymbol{x}^\star}$ is consistent. That is,*

$$\hat{\omega}_{\boldsymbol{x}^\star} - \omega_{\boldsymbol{x}^\star} \overset{P}{\to} 0, \quad as \quad n \to \infty.$$

Proof of Theorem A1. The proof is very straightforward. Lemma 2 shows that

$$\mu(t) \sim \int_{\mathcal{X}} \mathbb{E}[Y \mid T = t, \theta(\mathbf{x})] p_{\theta(\mathbf{X})}(\mathbf{x}) d\mathbf{x} \qquad \text{for } t \to \tau_R.$$

Assumption 4 allows us to rewrite (correctness of this step follows directly from Lemma A1 by considering $f(t, x) = \mathbb{E}[Y \mid T = t, \theta(\mathbf{x})]$ and $g(t, x) = \alpha(\theta(\mathbf{x})) + \beta(\theta(\mathbf{x}))t$)

$$\int_{\mathcal{X}} \mathbb{E}[Y \mid T = t, \theta(\mathbf{x})] p_{\theta(\mathbf{X})}(\mathbf{x}) d\mathbf{x} \sim \int_{\mathcal{X}} [\alpha(\theta(\mathbf{x})) + \beta(\theta(\mathbf{x}))t] \, p_{\theta(\mathbf{X})}(\mathbf{x}) d\mathbf{x} := \tilde{\mu}(t) \qquad \text{for } t \to \tau_R.$$

Since θ, α, and β are continuous and their estimators are consistent, we obtain that for all $t \in \mathcal{T}$, it holds that

$$\int_{\mathcal{X}} [\hat{\alpha}(\hat{\theta}(\mathbf{x})) + \hat{\beta}(\hat{\theta}(\mathbf{x}))t] \, p_{\theta(\mathbf{X})}(\mathbf{x}) d\mathbf{x} \overset{P}{\to} \tilde{\mu}(t) \quad \text{as } n \to \infty.$$

Moreover, from the law of large numbers, it holds that

$$\hat{\mu}(t) - \int_{\mathcal{X}} [\hat{\alpha}(\hat{\theta}(\mathbf{x})) + \hat{\beta}(\hat{\theta}(\mathbf{x}))t] \, p_{\theta(\mathbf{X})}(\mathbf{x}) d\mathbf{x} \overset{P}{\to} 0, \quad \text{as } n \to \infty.$$

Together, we obtain

$$\hat{\mu}(t) \overset{P}{\to} \tilde{\mu}(t), \quad \text{as } n \to \infty,$$

where the function on the right side is tail-equivalent with $\mu(t)$, which is what we wanted to show. $\square$

Lemma A1. *Let $\mathcal{X}$ be compact set and τ_R be the right endpoint of $\mathcal{T} \subseteq \mathbb{R}$. Let $f, g : \mathcal{T} \times \mathcal{X} \to \mathbb{R}$ be continuous functions such that for all $x \in \mathcal{X}$ holds $f(t,x) \sim g(t,x)$ as $t \to \tau_R$. Let F be a continuous distribution function. Then,*

$$\int_{\mathcal{X}} f(t,x) dF(x) \sim \int_{\mathcal{X}} g(t,x) dF(x), \quad as \ t \to \tau_R.$$

Proof. Let $\varepsilon > 0$. Find t_0 such that for all $t > t_0$ and for all $x \in \mathcal{X}$ holds $1 - \varepsilon < \frac{f(t,x)}{g(t,x)} < 1 + \varepsilon$. Then, for any $t > t_0$, it holds that

$$\frac{\int_{\mathcal{X}} f(t,x) dF(x)}{\int_{\mathcal{X}} g(t,x) dF(x)} < \frac{\int_{\mathcal{X}} f(t,x) dF(x)}{\int_{\mathcal{X}} \frac{1}{1+\varepsilon} f(t,x) dF(x)} = 1 + \varepsilon$$

and analogously $\frac{\int_{\mathcal{X}} f(t,x) dF(x)}{\int_{\mathcal{X}} g(t,x) dF(x)} > 1 - \varepsilon$. The proof is finished by sending $\varepsilon \to 0$. $\square$

Proof of Theorem A2. Idea: *We assume that the GPD approximation and linear model approximations are correct up to a factor of ε: we argue that for a large n this is correct. Next, we use Theorem 5.1 in [74] to show consistency of $\hat{\tau}$. We use Theorem 4.4 in [73] to show consistency of $\hat{\beta}$ together with linearity of the least square estimate (to show that it does not depend on the inaccuracy of the estimate of $\hat{\tau}$). Finally, we use Lemma 2 and send $\varepsilon \to 0$.*

Proof: Let $\varepsilon > 0$. We claim that it is possible to find $t < \tau_R$ and $n_0 \in \mathbb{N}$ such that for all $\mathbf{x} \in \mathcal{X}$ and all $n \geq n_0$, the following five statements hold with arbitrarily large probability:

- $t < F^{-1}_{T|\mathbf{X}=\mathbf{x}}(q_n)$ (In other words, q_{n_0} is large enough such that the q_n-quantile of $T \mid \mathbf{X}$ is larger than t.);

- It holds that

$$1 - \varepsilon < \frac{\mu_{\mathbf{x}^\star}(t)}{\mathbb{E}[Y \mid T = t, \tau(\mathbf{X}) = \tau(\mathbf{x}^\star)]} < 1 + \varepsilon,$$

where $\tau(\mathbf{x}) = \tau^\top_{lin} \mathbf{x}$ is the q_{n_0}-quantile of $T \mid \mathbf{X} = x$;

- It holds that

$$1 - \varepsilon < \frac{\mathbb{E}[Y \mid T = t, \tau(\mathbf{X}) = \tau(\mathbf{x}^\star)]}{\alpha \tau(\mathbf{x}^\star) + \beta \tau(\mathbf{x}^\star) t} < 1 + \varepsilon;$$

- $||\hat{\tau}_{lin} - \tau_{lin}|| < \varepsilon$;

- $|\hat{\beta} - \beta| < \varepsilon$;

where $\hat{\tau}_{lin} = argmin_b \sum_{i=1}^{n_0} h_{q_{n_0}}(T_i - \mathbf{X}^\top_i b)$ is the maximum likelihood estimator, and β is the real coefficient in the model

$$\mathbb{E}[Y \mid T = t, T > \tau^\top_{lin} \mathbf{x}, \mathbf{X} = \mathbf{x}] = \alpha \tau^\top_{lin} \mathbf{x} + \beta \tau^\top_{lin} \mathbf{x} t, \tag{A6}$$

and $\hat{\beta}$ is the corresponding least square estimate.

We prove the following bullet-points here:

- The first bullet-point is a trivial consequence of the assumption $q_n \to 1$.
- The second bullet-point is a trivial consequence of Lemma 2 together with Assumption D.
- The third bullet-point is a trivial consequence of Assumptions 4 and D;
- The fourth bullet-point follows from a well-known consistency of $\hat{\tau}_{lin}$. It is well known that for a fixed quantile q, the maximum likelihood estimator $\hat{\tau}_{lin} = argmin_b \sum_{i=1}^{n} h_q(T_i - \mathbf{X}^\top_i b)$ is consistent and even asymptotically normal (see, e.g., Theorem 4.1 in [68], noting that we assume continuous T and finite second moments of $\mathbf{X}$). However, quantile q is not fixed and is increasing with the sample size with the speed $\lim_{n \to \infty} q_n = 1$ and $\lim_{n \to \infty} n(1 - q_n) = \infty$. This is a well-known generalization of quantile regression known as 'intermediate order regression quantiles' or 'moderately extreme quan-

tiles' [75] and is as consistent and asymptotically normal under Assumption C (see Theorem 5.1 in [74]).

- The fifth bullet-point: For a moment, fix $\tau_{lin} \neq \mathbf{0}$. It is an elementary knowledge that the estimation of β using least squares in a model (A6), where τ_{lin} is fixed, consistent, and even asymptotically normal under conditions $var(Y) < \infty$, $\mathbb{E}||\mathbf{X}||^2 < \infty$, $(\mathbf{X}, T)$ satisfying Grenander conditions and the sample-size $|S| =: k_n = n(1 - q_n) \to \infty$ (see, e.g., Lemma A2). Observe that least squares estimate $\hat{\beta}$ is linear in τ_{lin}, that is, if we express $\hat{\beta}$ explicitly, we obtain $\hat{\beta} = \tau_{lin}^\top \tilde{\bar{\beta}}$, where $\tilde{\beta}$ is a coefficient in a linear model corresponding to (A9) (where T is assumed to be larger than $\tau^\top X$ implicitly). Finally, using this observation, we can replace the fixed value of τ_{lin} by a random $\hat{\tau}_{lin}$, and we still obtain $\hat{\beta} = \hat{\tau}_{lin}^\top \tilde{\bar{\beta}}$. Since by increasing n we can make $\tilde{\bar{\beta}}$ arbitrarily accurate with arbitrarily large probability, the same holds for $\hat{\beta}$. In the following paragraph, we present an an illustration of the linearity of $\hat{\beta}$ in τ_{lin} for $d = 1$. An explicit expression of $\hat{\beta}$ as a function of τ_{lin} and our data is the following:

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = (M^\top M)^{-1} M^\top \mathbf{Y}_S, \quad where \quad M = \begin{bmatrix} \tau_{lin}\mathbf{x}_1 & \tau_{lin}\mathbf{x}_1 t_1 \\ \dots & \dots \\ \tau_{lin}\mathbf{x}_k & \tau_{lin}\mathbf{x}_k t_k \end{bmatrix} \tag{A7}$$

where $\mathbf{Y}_S = (Y_1, \dots, Y_k)^\top$, WLOG let $S = \{1, \dots, k_n\} \subset \{1, \dots, n\}$. Note that

$$M = \tau_{lin} diag(x_1, \dots, x_k) \begin{bmatrix} 1 & t_1 \\ \dots & \dots \\ 1 & t_k \end{bmatrix} = \tau_{lin}\tilde{M}, \tag{A8}$$

where $\tilde{M}$ is the data matrix corresponding to a model (A9).

Combining all the bullet-points, we obtain that with arbitrarily large probability that

$$\begin{aligned} \mu_{\mathbf{x}^\star}(t) &\approx \mathbb{E}[Y \mid T = t, \tau(\mathbf{X}) = \tau(\mathbf{x})] \\ &\approx \alpha\tau(\mathbf{x}) + \beta\tau(\mathbf{x})t \\ &\approx \hat{\alpha}\hat{\tau}_{lin}^\top \mathbf{x}^\star + \hat{\beta}\hat{\tau}_{lin}^\top \mathbf{x}^\star t, \end{aligned}$$

where each sign '$\approx$' represent equality up to a factor of ε (in either multiplicative or additive form) which is negligible as $\varepsilon \to 0$. This implies consistency, Quod erat demonstrandum. $\square$

Lemma A2. *Consider an estimate* $(\hat{\alpha}, \tilde{\bar{\beta}})$ *of* $\alpha \in \mathbb{R}, \tilde{\beta} \in \mathbb{R}^d$ *using least squares in a model*

$$\mathbb{E}[Y \mid T = t, \mathbf{X} = \mathbf{x}] = \alpha\mathbf{x} + \tilde{\beta}^\top \mathbf{x}t, \tag{A9}$$

based on a random sample $(Y_1, T_1, X_1), \dots (Y_k, T_k, X_k)$. *Then,* $\tilde{\bar{\beta}}$ *is consistent and asymptotically normal if* $var(Y) < \infty$ *and* $\mathbb{E}||\mathbf{X}||^2 < \infty$ *and* $(\mathbf{X}, T)$ *satisfy Grenander conditions.*

Proof can be found in Theorem 4.4 in [73].

Appendix D.4. Bootstraps Correctness

We use results from Chapter 23.2 in [58]. The main step is to use the delta method for bootstrap (Theorem 23.5 in [58]) and the fact that regression models in step 1 and step 2 of our algorithm are 'bootstrappable' (Theorem 3 in [76,77]).

To simplify some steps of the proof, we assume the following:

E. Assume that $\hat{\omega}_{\mathbf{x}^\star}$ is consistent (which holds for example under assumptions A,B,C,D).

F. We compute $\hat{\tau}_{lin}$ from the first $\lfloor \frac{n}{2} \rfloor$ data points, and we compute $\hat{\beta}$ from the remaining $\lceil \frac{n}{2} \rceil$ data points.

G. In the computation of the set S, we assume that τ is known and non-random; that is, $S = \{i \leq n : T_i > \tau(\mathbf{X}_i)\}$ instead of $S = \{i \leq n : T_i > \hat{\tau}(\mathbf{X}_i)\}$.

H. Assumption of Theorem 3 in [76] are satisfied; that is, $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ is non-singular matrix, the conditional density of $Y - \tau^\top\mathbf{X}$ given $\mathbf{X}$, denoted as f, satisfies $f(\epsilon \mid \mathbf{X}) > r_1$ whenever $|\epsilon| \leq r_2$ for some positive numbers r_1, r_2. Finally, there exists some function G such that $f(\epsilon \mid \mathbf{X}) \leq G(\mathbf{X})$ for all ϵ and $\mathbb{E}\big[(1 + G(\mathbf{X}))||\mathbf{X}||^2\big] < \infty$.

Theorem A3. *Assume validity of assumptions D,E,F,G and H. Let $q \in (0,1)$ be chosen such that the distribution of $T \mid T > \tau_q(\mathbf{x}), \mathbf{X} = \mathbf{x}$ follows $GPD(\theta(\mathbf{x}))$ for all $\mathbf{x} \in \mathcal{X}$. Then, $\hat{\zeta}_\alpha$ is asymptotically consistent; that is,*

$$liminf_{n\to\infty} P(\omega_{\mathbf{x}^\star} \leq \hat{\zeta}_\alpha) \geq 1 - \alpha.$$

Proof. We will show that both $\frac{1}{\sqrt{n}}(\hat{\omega}_{\mathbf{x}^\star} - \omega_{\mathbf{x}^\star})$ and $\frac{1}{\sqrt{n}}(\hat{\omega}^\star_{\mathbf{x}^\star} - \hat{\omega}_{\mathbf{x}^\star})$ given P_n both converge to the same distribution, say G. That is, $\frac{1}{\sqrt{n}}(\hat{\omega}_{\mathbf{x}^\star} - \omega_{\mathbf{x}^\star}) \xrightarrow{D} G \xleftarrow{P} \frac{1}{\sqrt{n}}(\hat{\omega}^\star_{\mathbf{x}^\star} - \hat{\omega}_{\mathbf{x}^\star})$ as $n \to \infty$. This directly implies (see, e.g., Lemma 23.3 in [58]) that $\hat{\zeta}_\alpha$ is asymptotically consistent.

(Observation 1) $\hat{\tau}_{lin}$ satisfies that $\frac{1}{\sqrt{n}}(\hat{\tau}_{lin} - \tau_{lin})$ and $\frac{1}{\sqrt{n}}(\hat{\tau}^\star_{lin} - \hat{\tau}_{lin})$ given P_n both converge to the same Gaussian distribution (see Theorem 3 in [76]).

(Observation 2) $\hat{\beta} = \hat{\tau}_{lin}^\top\widehat{\widetilde{\beta}}$, where $\tilde{\beta}$ is a coefficient in a linear model corresponding to (A9) (where T is assumed to be larger than $\tau(\mathbf{X})$ implicitly since we assumed that τ is known and non-random in S). Note that $\widehat{\widetilde{\beta}} \perp\!\!\!\perp \hat{\tau}_{lin}$ since $\widehat{\widetilde{\beta}}$ is computed from the second half of the dataset and its computation does not contain $\hat{\tau}_{lin}$. However, we know that $\widehat{\widetilde{\beta}}$ satisfies that $\frac{1}{\sqrt{n}}(\widehat{\widetilde{\beta}} - \tilde{\beta})$ and $\frac{1}{\sqrt{n}}(\widehat{\widetilde{\beta}}^\star - \widehat{\widetilde{\beta}})$ given P_n both converge to the same Gaussian distribution (Theorem 2 in [78] or [77]).

Together: Since $\hat{\tau}_{lin}$ is 'bootstrappable' and $\widehat{\widetilde{\beta}}$ is 'bootstrappable' and they are independent, the delta method give us that $\hat{\omega}_{\mathbf{x}^\star}$ is 'bootstrappable'. More formally, we use Theorem 23.5 in [58] (Delta method for bootstrap). Define $\phi : \mathbb{R}^{2d} \to \mathbb{R} : \phi(a,b) = (a^\top b)(a^\top\mathbf{x}^\star)$. Note that $\hat{\omega}_{\mathbf{x}^\star} = \phi(\hat{\tau}_{lin}, \widehat{\widetilde{\beta}})$. Since $\hat{\tau}_{lin}$ and $\widehat{\widetilde{\beta}}$ satisfy the conditions of the theorem, we obtain that $\frac{1}{\sqrt{n}}(\hat{\omega}_{\mathbf{x}^\star} - \omega_{\mathbf{x}^\star})$ and $\frac{1}{\sqrt{n}}(\hat{\omega}^\star_{\mathbf{x}^\star} - \hat{\omega}_{\mathbf{x}^\star})$ given P_n both converge to the same distribution. That is what we wanted to show. $\square$

Appendix E. Proofs of Lemmas 1 and 2

Proof of Lemma 1. A simple computation gives us

$$\mathbb{E}[Y(t)] = \mathbb{E}(\mathbb{E}[Y(t) \mid \mathbf{X}]) \sim \mathbb{E}(\mathbb{E}[Y \mid \mathbf{X}, T = t])$$
$$= \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{Y|\mathbf{X},T}(y, \mathbf{x}, t) p_{\mathbf{X}}(\mathbf{x}) y \ dy d\mathbf{x}$$
$$= \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{p_T(t)}{p_{T|\mathbf{X}}(t \mid \mathbf{x})} p_{Y,\mathbf{X}|T}(y, \mathbf{x} \mid t) y \ dy d\mathbf{x}$$
$$= \mathbb{E}\{\pi_0(T, \mathbf{X}) Y \mid T = t\}.$$

$\square$

Proof of Lemma 2. From Assumption 2, we have that

$$\mathbb{E}[Y(t) \mid \mathbf{X}] \sim \mathbb{E}[Y \mid T = t, \mathbf{X}] \qquad \text{as } t \to \tau_R. \tag{A10}$$

On both sides of (A10), we condition on $\theta(\mathbf{X})$ and integrate over the remaining $\mathbf{X}$ (denoted as $\theta^C(\mathbf{X})$, formally it is an orthogonal complement). Note that the distribution of $T \mid \theta(\mathbf{X})$ approaches the distribution of $T \mid \mathbf{X}$, given $T > \tau(\mathbf{X})$ for sufficiently large $\tau(\mathbf{X})$, since it

approaches $GPD(\theta(\mathbf{X}))$. Hence, $P_{\theta^C(\mathbf{X})|T=t,\theta(\mathbf{X})}$ approaches the distribution $P_{\theta^C(\mathbf{X})|\theta(\mathbf{X})}$ as $t \to \tau_R$.

We obtain the following:

$$
\begin{aligned}
\mathbb{E}[Y(t) \mid \theta(\mathbf{X})] &= \int \mathbb{E}[Y(t) \mid \theta(\mathbf{X}), \theta^C(\mathbf{X}) = w] dP_{\theta^C(\mathbf{X})|\theta(\mathbf{X})}(w) \\
&\sim \int \mathbb{E}[Y(t) \mid \theta(\mathbf{X}), \theta^C(\mathbf{X}) = w] dP_{\theta^C(\mathbf{X})|T=t,\theta(\mathbf{X})}(w) \\
&= \int \mathbb{E}[Y \mid T = t, \theta(\mathbf{X}), \theta^C(\mathbf{X}) = w] dP_{\theta^C(\mathbf{X})|T=t,\theta(\mathbf{X})}(w) \\
&= \mathbb{E}[Y \mid T = t, \theta(\mathbf{X})].
\end{aligned}
$$

The second statement in the Lemma trivially follows from the first by integrating over $\theta(\mathbf{X})$. $\square$

References

1. Rosenbaum, P.R.; Rubin, D.B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
2. Holland, P.W. Statistics and Causal Inference. *J. Am. Stat. Assoc.* **1986**, *81*, 945–960. [CrossRef]
3. Robins, J.M.; Hernández-Díaz, S.; Brumback, B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* **2000**, *11*, 550–560. [CrossRef] [PubMed]
4. Imai, K.; King, G.; Stuart, E.A. Misunderstandings Between Experimentalists and Observationalists about Causal Inference. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2008**, *171*, 481–502. [CrossRef]
5. Imai, K.; van Dyk, D.A. Causal Inference With General Treatment Regimes. *J. Am. Stat. Assoc.* **2004**, *99*, 854–866. [CrossRef]
6. Heckman, J.J.; Humphries, J.E.; Veramendi, G. Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking. *J. Political Econ.* **2018**, *126*, 197–246. [CrossRef] [PubMed]
7. Hannart, A.; Naveau, P. Probabilities of Causation of Climate Changes. *J. Clim.* **2018**, *31*, 5507–5524. [CrossRef]
8. Low, H.; Meghir, C. The Use of Structural Models in Econometrics. *J. Econ. Perspect.* **2017**, *31*, 33–58. [CrossRef]
9. Rubin, D.B. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *J. Am. Stat. Assoc.* **2005**, *100*, 322–331. [CrossRef]
10. Imbens, G.W.; Rubin, D.B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*; Cambridge University Press: Cambridge, UK, 2015. [CrossRef]
11. Kennedy, E.H.; Ma, Z.; McHugh, M.D.; Small, D.S. Non-parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2017**, *79*, 1229–1245. [CrossRef]
12. Westling, T.; Gilbert, P.; Carone, M. Causal Isotonic Regression. *JRSSb* **2020**, *82*, 719–747. [CrossRef]
13. Galagate, D. Causal Inference with a Continuous Treatment and Outcome: Alternative Estimators for Parametric Dose-Response Functions with Applications. Ph.D. Thesis, University of Maryland, College Park, MD, USA, 2016. [CrossRef]
14. Rubin, D.; van der Laan, M.J. *Extending Marginal Structural Models through Local, Penalized, and Additive Learning*; Working Paper 212; Division of Biostatistics, UC Berkeley: Berkeley, CA, USA, 2006.
15. Neugebauer, R.; van der Laan, M.J. Nonparametric causal effects based on marginal structural models. *J. Stat. Plan. Inference* **2007**, *137*, 419–434. [CrossRef]
16. Zhang, Y.F.; Zhang, H.; Lipton, C.Z.; Li, L.E.; Xing, E. Exploring Transformer Backbones for Heterogeneous Treatment Effect Estimation. *NeurIPS ML Safety Workshop* **2023**. [CrossRef]
17. Bica, I.; Jordon, J.v.d.; Schaar, M. Estimating the Effects of Continuous-valued Interventions using Generative Adversarial Networks. In Proceedings of the Advances in Neural Information Processing Systems, virtual, 6–12 December 2020; Volume 33, pp. 16434–16445. [CrossRef]
18. Zhang, Y. Extremal Quantile Treatment Effects. *Ann. Stat.* **2018**, *46*, 3707–3740. [CrossRef]
19. Deuber, D.; Li, J.; Engelke, S.; Maathuis, M. Estimation and Inference of Extremal Quantile Treatment Effects for Heavy-Tailed Distributions. *JASA* **2023**, 1–11. [CrossRef]
20. Huang, W.; Li, S.; Peng, L. Extreme Continuous Treatment Effects: Measures, Estimation and Inference. *arXiv* **2022**, arXiv:2209.00246.
21. Bodik, J.; Paluš, M.; Pawlas, Z. Causality in extremes of time series. *Extremes* **2024**, *27*, 67–121. [CrossRef]
22. Gnecco, N.; Meinshausen, N.; Peters, J.; Engelke, S. Causal discovery in heavy-tailed models. *Ann. Stat.* **2020**, *49*, 1755–1778. [CrossRef]
23. Pasche, O.C.; Chavez-Demoulin, V.; Davison, A. Causal Modelling of Heavy-Tailed Variables and Confounders with Application to River Flow. *Extremes* **2023**, *26*, 573–594. [CrossRef]
24. Krali, M.; Davison, A.C.; Klüppelberg, C. Heavy-tailed max-linear structural equation models in networks with hidden nodes. *arXiv* **2023**, arXiv:2306.15356.

25. Bodik, J.; Chavez-Demoulin, V. Structural restrictions in local causal discovery: Identifying direct causes of a target variable. *arXiv* **2023**, arXiv:2307.16048.
26. Engelke, S.; Hitz, A. Graphical models for extremes. *J. R. Stat. Soc. Ser. B* **2020**, *82*, 871–932. [CrossRef]
27. Naveau, P.; Hannart, A.; Ribes, A. Statistical Methods for Extreme Event Attribution in Climate Science. *Annu. Rev. Stat. Its Appl.* **2020**, *7*, 89–110. [CrossRef]
28. Courgeau, V.; Veraart, A.E.D. Extreme event propagation using counterfactual theory and vine copulas. *arXiv* **2021**, arXiv:2106.13564.
29. Kiriliouk, A.; Naveau, P. Climate extreme event attribution using multivariate peaks-over-thresholds modeling and counterfactual theory. *Ann. Appl. Stat.* **2020**, *14*, 1342–1358. [CrossRef]
30. Dong, K.; Ma, T. First steps toward understanding the extrapolation of nonlinear models to unseen domains. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
31. Christiansen, R.; Pfister, N.; Jakobsen, M.E.; Gnecco, N.; Peters, J. A Causal Framework for Distribution Generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6614–6630. [CrossRef]
32. Saengkyongam, S.; Rosenfeld, E.; Ravikumar, P.K.; Pfister, N.; Peters, J. Identifying Representations for Intervention Extrapolation. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
33. Chen, Y.; Bühlmann, P. Domain Adaptation under Structural Causal Models. *J. Mach. Learn. Res.* **2021**, *22*, 1–80.
34. Shen, X.; Meinshausen, N. Engression: Extrapolation for Nonlinear Regression? *arXiv* **2024**, arXiv:2307.00835.
35. Pfister, N.; Bühlmann, P. Extrapolation-Aware Nonparametric Statistical Inference. *arXiv* **2024**, arXiv:2402.09758
36. Hirano, K.; Imbens, G.W. The Propensity Score with Continuous Treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*; John Wiley and Sons, Ltd.: Hoboken, NJ, USA, 2004; Chapter 7, pp. 73–84. [CrossRef]
37. Gill, R.D.; Robins, J.M. Causal inference for complex longitudinal data: The continuous case. *Ann. Stat.* **2001**, *29*, 1785–1811. [CrossRef]
38. King, G.; Zeng, L. The Dangers of Extreme Counterfactuals. *Political Anal.* **2006**, *14*, 131–159. [CrossRef]
39. Crump, R.K.; Hotz, V.J.; Imbens, G.W.; Mitnik, O.A. Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika* **2009**, *96*, 187–199. [CrossRef]
40. Ai, C.; Linton, O.; Zhang, Z. Estimation and Inference for the Counterfactual Distribution and Quantile Functions in Continuous Treatment Models. *J. Econom.* **2021**, *228*, 39–61. [CrossRef]
41. Bahadori, M.T.; Tchetgen, E.; Heckerman, D. End-to-End Balancing for Causal Continuous Treatment-Effect Estimation. In Proceedings of the 39th International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 1313–1326.
42. Li, Y.; Kuang, K.; Li, B.; Cui, P.; Tao, J.; Yang, H.; Wu, F. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding. In Proceedings of the 2020 KDD Workshop on Causal Discovery, PMLR, San Diego, CA, USA, 24 August 2020; Volume 127, pp. 4–22.
43. Kreif, N.; Grieve, R.; Díaz, I.; Harrison, D. Evaluation of the Effect of a Continuous Treatment: A Machine Learning Approach with an Application to Treatment for Traumatic Brain Injury. *Health Econ.* **2015**, *24*, 1213–1228. [CrossRef] [PubMed]
44. Zhao, S.; van Dyk, D.A.; Imai, K. Propensity Score-based Methods for Causal Inference in Observational Studies with Non-binary Treatments. *Stat. Methods Med. Res.* **2020**, *29*, 709–727. [CrossRef] [PubMed]
45. Resnick, S.I. *Extreme Values, Regular Variation and Point Processes*; Springer: New York, NY, USA, 2008.
46. Pickands, J. Statistical Inference Using Extreme Order Statistics. *Ann. Stat.* **1975**, *3*, 119–131. [CrossRef]
47. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*; Springer Series in Statistics; Springer: London, UK, 2001.
48. Fisher, R.; Tippett, L. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math. Proc. Camb. Philos. Soc.* **1928**, *24*, 180–190. [CrossRef]
49. Pearl, J. *Causality: Models, Reasoning and Inference*; Cambridge University Press: Cambridge, UK, 2009.
50. Soklakov, A. Occam's razor as a formal basis for a physical theory. *Found. Phys. Lett.* **2002**, *15*, 107–135. [CrossRef]
51. Wood, S. *Generalized Additive Models: An Introduction with R*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2017.
52. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
53. Smith, R. Extreme Value Theory. *Handb. Appl. Math.* **1990**, *1*, 437–447.
54. Davison, A.; Huser, R. Statistics of Extremes. *Annu. Rev. Stat. Its Appl.* **2015**, *2*, 203–235. [CrossRef]
55. Schneider, L.F.; Krajina, A.; Krivobokova, T. Threshold selection in univariate extreme value analysis. *Extremes* **2021**, *24*, 881–913. [CrossRef]
56. Caeiro, F.; Gomes, M. Threshold selection in extreme value analysis. In *Extreme Value Modeling and Risk Analysis: Methods and Applications*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2015; pp. 69–82, ISBN 9780429161193.
57. Davison, A.; Smith, R.L. Models for exceedances over high thresholds. *J. R. Stat. Soc. Ser. B (Methodol.)* **1990**, *52*, 393–425. [CrossRef]
58. van der Vaart, A.W. *Bootstrap*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge, UK, 1998; pp. 326–340.
59. Kennedy, E.H. Nonparametric causal effects based on incremental propensity score interventions. *JASA* **2019**, *114*, 645–656. [CrossRef]
60. Bia, M.; Mattei, A.; Nicolò, G. A Stata package for the application of semiparametric estimators of dose response functions. *Stata J.* **2014**, *14*, 580–604. [CrossRef]

61. Galagate, D.; Schafer, J. *Estimating Causal Dose Response Functions*; R package version 0.4.2; R Foundation for Statistical Computing: Vienna, Austria, 2022.
62. van der Wal, W.; Geskus, R. IPW: An R package for inverse probability weighting. *J. Stat. Softw.* **2011**, *43*, 1–23.
63. Davis, R.A. The rate of convergence in distribution of the maxima. *Stat. Neerl.* **1982**, *36*, 31–35. [CrossRef]
64. Engelke, S.; Ivanovs, J. Sparse structures for multivariate extremes. *Annu. Rev. Stat. Its Appl.* **2021**, *8*, 241–270. [CrossRef]
65. Pearl, J. Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, USA, 2–5 August 2001; UAI'01; pp. 411–420.
66. Yeh, I. Modeling of strength of high performance concrete using artificial neural networks. *Cem. Concr. Res.* **1998**, *28*, 1797–1808. [CrossRef]
67. Neville, A. *Properties of Concrete*; Pearson Education Limited: London, UK, 2011.
68. Koenker, R. *Quantile Regression*; Econometric Society Monographs; Cambridge University Press: Cambridge, UK, 2005. [CrossRef]
69. Youngman, B.D. Evgam: An R Package for Generalized Additive Extreme Value Models. *J. Stat. Softw.* **2022**, *103*, 1–26. [CrossRef]
70. Kolesárová, A.; Mesiar, R.; Saminger-Platz, S. Generalized Farlie-Gumbel-Morgenstern Copulas. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*; Springer International Publishing: Cham, Switzerland, 2018; Communications in Computer and Information Science; Volume 853, pp. 244–252. [CrossRef]
71. Perlin, K. An Image Synthesizer. *SIGGRAPH Comput. Graph.* **1985**, *19*, 287–296. [CrossRef]
72. Davis, R.A.; Mikosch, T. The extremogram: A correlogram for extreme events. *Bernoulli* **2009**, *15*, 977–1009. [CrossRef]
73. Greene, W.H. *Econometric Analysis*; Pearson Education: London, UK, 2008.
74. Chernozhukov, V. Extremal quantile regression. *Ann. Stat.* **2005**, *33*, 806–839. [CrossRef]
75. Chernozhukov, V.; Fernández-Val, I.; Kaji, T. *Extremal Quantile Regression: An Overview*; Chapman and Hall: Boca Raton, FL, USA, 2016.
76. Hahn, J. Bootstrapping Quantile Regression Estimators. *Econom. Theory* **1995**, *11*, 105–121. [CrossRef]
77. Freedman, D.A. Bootstrapping Regression Models. *Ann. Stat.* **1981**, *9*, 1218–1228. [CrossRef]
78. Eck, D.J. Bootstrapping for multivariate linear regression models. *Stat. Probab. Lett.* **2018**, *134*, 141–149. [CrossRef]