Causal Inference in the Tails: Extreme Event Attribution, Treatment Effects, and Structural Discovery

Mengran Li

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy

School of Mathematics & Statistics

College of Science & Engineering

May 2026

Abstract

Many consequential questions about rare, high-impact events, including heatwaves, financial crashes, severe adverse drug responses, and extreme precipitation, arise in the tail of the distribution. These events are difficult to analyse because observations are scarce, methods designed for typical outcomes can be unreliable in the regions of interest, and the questions being asked often concern interventions, counterfactuals, or structural propagation rather than descriptive probabilities alone. Extreme value theory provides principled tools for modelling tail behaviour, while causal inference provides a language for interventions and counterfactual reasoning. Yet standard causal assumptions about support, overlap, and stability are often most fragile in the tail. This thesis develops a coherent methodological perspective on causal inference for extremes through the connected tasks of attribution, estimation, and discovery.

The first contribution addresses attribution through a sensitivity framework within a counterfactual setting, examining how attribution metrics vary across alternative marginal and joint-tail specifications. In the simulations and two precipitation applications, estimates vary materially across the three fitted tail models, in some cases changing sign. The results show that attribution conclusions can be sensitive to both marginal and dependence modelling choices. The second contribution focuses on estimation via the Tail-Calibrated Inverse Estimating Equation (TIEE) for extreme quantile treatment effects. TIEE combines causal adjustment with tail calibration, allowing information from less extreme quantiles to stabilise inference at extreme levels where direct quantile estimation is unreliable. By embedding extreme value structure within a unified estimating equation, TIEE enables rigorous causal attribution for outcomes so rare that standard methods become ineffective. The third contribution addresses discovery through Sparse Structure diScovery in Multivariate Extremes (S3ME), a two-stage method for learning sparse ex-

tremal causal structure. By combining skeleton estimation with edge orientation based on tail-induced asymmetry, S3ME recovers propagation pathways in settings characterised by heavy tails and latent shocks, with applications to hydrological and financial tail-risk systems.

Across these three components, the thesis shows that causal inference in the tails cannot be treated as a straightforward extension of standard methods. Tail assumptions fundamentally shape the estimands, the stability of inference, and the recoverable structures. As such, extreme value modelling and causal reasoning must be developed jointly, rather than applied as separate sequential tools.

Contents

List of Tables

List of Figures

List of symbols

Symbol	Meaning
X, Y	Observed random variables or outcome variables; chapter-specific definitions are given where first used.
$X^{(t)}$, $Y(t)$	Potential variables or outcomes under treatment/world $t \in \{0, 1\}$.
p_t	Probability of the target extreme event under world t.
v	Event-defining high threshold.
u, $\boldsymbol{u}$, $\boldsymbol{u}^{(t)}$	Scalar or componentwise intermediate thresholds used to define the tail region; $\boldsymbol{u}^{(t)}$ is the threshold vector in world t.
$\boldsymbol{w}$	Nonnegative projection-weight vector for a multivariate extreme event.
τ	Target quantile level.
$\theta_{d,\tau}$	Quantile parameter for treatment state d at level τ.
$s_d(W, \theta, \zeta)$	Signal function identifying the treatment-specific distribution or quantile parameter.
$\pi_d(\boldsymbol{X}; \zeta)$	Propensity score for treatment state d.
α_j	Positive rate parameter of the jth independent reverse-exponential generator component in the extreme-event-attribution chapter; α_1 and α_2 denote the two bivariate component rates.
ϵ_j	Innovation or noise variable for node j in Chapter 6.
$\mathcal{G} = (V, E)$	Directed acyclic graph with node set V and edge set E.
$\mathrm{pa}(j)$	Parent set of node j.
P	Observed proxy for latent common shocks.
k	Number of observations retained in the tail region.
$d_{\max}$	Maximum permitted in-degree in the orientation search.
$s_{\max}$	Maximum degree of the true undirected skeleton.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Daniela Castro-Camilo. We have been working together since a period that was personally very difficult for me. Throughout this time, she has always been warm, encouraging, and supportive. I am profoundly grateful to her for introducing me to the field of extreme value theory, exposing me to the area of causality, and providing me with exceptional PhD training. I have learned enormously from her, from how to think critically to how to write clearly and effectively. She has always cared deeply about my professional development, helping me secure opportunities to present at conferences, supporting me in competitive endeavors, and opening doors to a much wider academic world.

I would also like to thank the School of Mathematics and Statistics at the University of Glasgow for providing such a supportive and welcoming research environment. I am grateful to the academic, administrative, and support staff for their kindness and help throughout my PhD. The warmth of this community made the School a place where I felt encouraged, included, and at home even while living far away from home.

I am also grateful to the EVT community, and especially to everyone involved in the Glasgow–Edinburgh Extremes Network (GLE^{2}N). The monthly discussions have been one of the most valuable parts of my PhD experience, offering not only thoughtful feedback and new perspectives, but also a strong sense of connection to a wider community of researchers working on extremes.

Finally, I would like to thank my family and friends for their unwavering support and encouragement. My parents, Xinshui Li and Xianzhi Zuo, have always believed in me and supported my dreams, even when they were far removed from their own experiences. They have been a constant source of love and encouragement throughout my life. I am also grateful to my friends for their companionship, understanding, and support during this journey.

I would like to dedicate this thesis to the memory of my father.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Chapter 1
Introduction

1.1 Motivation

Many of the most consequential questions about rare, high-impact events arise in the tail of the distribution. In climate science, an extreme heatwave or flood that occurs once in decades can have lasting consequences for societies and infrastructure, yet such events remain difficult to attribute, anticipate, and analyse with standard inferential tools. Similar difficulties appear in finance and medicine, where crises and severe adverse responses are rare but disproportionately important. In each of these settings, observations are sparse, dependence in the tail can differ markedly from that in the bulk, and methods designed for typical outcomes may be unreliable precisely where the stakes are highest (Resnick, 2007; Beirlant et al., 2004).

Descriptive tools alone are insufficient for studying such events. They can reveal how often extremes occur, how severe they are, and which variables tend to co-move in the tail. Statistical models of extremes go further, fitting tail distributions and computing return levels, but they too remain associational, describing the data-generating process as observed rather than how it would respond to intervention. The questions that matter most, however, are causal in nature. One may ask whether human activities have altered the probability of an extreme event, or whether an intervention would reduce the risk of a catastrophic outcome. When extremes propagate through a system, one wants to know which variables are drivers and which are consequences. None of these questions can be answered by characterising the tail alone, as they demand counterfactual reasoning and a structural understanding of how extremes are generated (Pearl, 2009b; Hernán and Robins, 2020).

Once the goal shifts from describing extremes to reasoning about them causally, three tasks come into view. The first is attribution, a retrospective task that asks, given an extreme event that has already occurred, to what extent an external driver such as human activity has altered its probability or severity. The second is estimation, a prospective task that asks how treatments or interventions change extreme outcomes, quantifying the effects of causes rather than the causes of effects. Attribution depends fundamentally on estimation, since determining whether a driver changed the probability of an observed extreme requires knowledge of the counterfactual distribution of outcomes under different levels of that driver. The third task is discovery, which involves recovering the structural relationships through which extremes propagate across a system. Such structural knowledge can in turn inform and strengthen estimation, so that the three tasks, though conceptually distinct, form a natural progression: attribution depends on estimation, and estimation can be informed by discovery, even if it does not strictly require it.

Motivated by this natural hierarchy, this thesis develops a coherent methodological perspective on causal extremes, providing both the theoretical foundations and the practical tools needed to address these three tasks in real-world settings.

1.2 Research gaps

The literature does not yet provide a unified account of causal inference for extreme outcomes. Extreme value theory offers principled tools for describing tail behaviour (Coles, 2001; Beirlant et al., 2004), but these tools are not designed to reason about interventions, counterfactuals, or causal structure. Causal inference, by contrast, provides formal frameworks for identification and intervention (Pearl, 2009b; Hernán and Robins, 2020), but these frameworks are usually developed under support, stability, and regularity conditions that are routinely satisfied in the bulk but often violated in the tail. Existing work has therefore made progress on the two sides of the problem largely in parallel rather than in combination.

This missing integration becomes visible across the three inferential tasks. In attribution, the validity of causal conclusions about rare events rests critically on assumptions about joint tail dependence, yet dependence uncertainty is rarely elevated to a central inferential concern (Naveau et al., 2020a; van Oldenborgh et al., 2021). In the estimation of extreme quantile effects, standard causal estimators and extreme value tools do not combine naturally into a coherent account of identification, adjustment, and extrapolation (Chernozhukov, 2005; Bhuyan et al., 2023; Deuber et al., 2024). In causal discovery, co-exceedance in multivariate extremes may reflect propagation, latent common shocks, or shared tail heaviness, while existing methods struggle to distinguish these possibilities in a principled way (Gnecco et al., 2021a; Engelke et al., 2025). These are not three arbitrary applications, but three tasks within the same broader problem.

The central gap addressed by this thesis is the absence of a unified account in which attribution, estimation, and discovery can be understood as connected components of causal inference beyond the bulk.

1.3 Research questions and contributions

This thesis poses three research questions, one for each of the inferential tasks introduced above, and develops a concrete methodological contribution in response to each. They follow the same progression: from attribution, to estimation, to discovery. The questions are as follows.

How sensitive are attribution conclusions to extremal dependence assumptions? Attribution metrics such as the attribution ratio and the probability of necessary causation depend not only on marginal tail behaviour but also on how joint extremes are modelled. Chapter 4 compares three parametric tail models through simulations and two precipitation applications. The estimated metrics vary materially across fitted tail specifications, motivating sensitivity analysis for both marginal and dependence assumptions.

How can treatment effects on extreme quantiles be estimated reliably? Standard quantile-based causal estimators become unstable at extreme levels due to data sparsity and the unreliability of pointwise inversion far into the tail. This thesis proposes replacing such inversion with an integrated estimating equation that combines causal adjustment with tail calibration, and establishes identification under unconfoundedness together with a tail overlap condition. The resulting estimator, TIEE, is shown to recover causal effects at extreme quantile levels where existing methods break down.

Can causal structure be recovered from tail behaviour? Bulk-based causal discovery methods rely on conditional independence in the centre of the distribution, where information about directional relationships in the tail may be absent or misleading. This thesis shows that, under recursive extremal structural equation models, tail asymmetry carries directional information that bulk methods cannot exploit, and proposes a two-stage procedure that first screens for a sparse candidate skeleton and then orients edges using this asymmetry. Identifiability and consistency results are established, and the approach is shown to recover causal structure in settings where bulk-based methods are poorly matched to the problem.

Taken together, these three contributions show that attribution, estimation, and discovery are not separate tail problems but interconnected parts of a broader programme of causal inference in extremes. The thesis contribution is therefore both methodological and conceptual, developing new tools for each task while making the case that causal reasoning beyond the bulk cannot be adequately pursued without engaging all three problems together.

1.4 Thesis outline

The remainder of this thesis is organised as follows. Chapter 2 introduces the theoretical background required for the subsequent chapters. It reviews the potential outcomes framework for causal inference and the fundamentals of univariate and multivariate extreme value theory, and examines why standard causal assumptions become fragile in the tail regime. The chapter establishes notation and defines the key objects that recur throughout the thesis.

Chapter 3 provides a critical review of the existing literature at the intersection of causal inference and extreme value theory. It surveys methods for extreme event attribution, approaches to quantile treatment effect estimation in the tails, and techniques for causal discovery under extremal dependence, situating the specific gaps that each subsequent chapter addresses.

Chapter 4 studies extreme event attribution under tail dependence uncertainty. It develops a systematic comparison of attribution metrics under alternative extremal dependence specifications, presents simulation experiments quantifying the sensitivity of attribution conclusions, and applies the framework to European and North American precipitation extremes.

Chapter 5 develops the Tail-Calibrated Inverse Estimating Equation (TIEE) framework for extreme quantile treatment effects. It introduces the integrated estimating equation, establishes theoretical properties, and evaluates performance through simulation studies and an application to extreme precipitation in the Austrian Alps, illustrating how TIEE enables observational causal attribution for very rare events under anthropogenic warming.

Chapter 6 addresses causal discovery in multivariate extremes. It develops methods for skeleton screening and edge orientation based on tail asymmetry, establishes high-dimensional sure screening and oracle consistency for a direct-fit global orientation score, and applies the framework to river network and financial data.

Chapter 7 synthesises the contributions, reflects on the coherent methodological perspective developed across the thesis, and discusses limitations and directions for future research.

Chapter 2

Theoretical background

This chapter motivates and sets up the theoretical tools used to study causal questions about rare and extreme outcomes. Such questions combine two difficulties. They are causal because they compare outcomes under alternative interventions or counterfactual worlds, only one of which is observed. They are extremal because the outcomes of interest lie in the tail of the distribution, where data are sparse and ordinary central approximations give little guidance. The chapter therefore introduces causal inference and extreme value theory as complementary languages for defining, identifying, and estimating the quantities studied later in the thesis. Section 2.1 introduces the counterfactual, identification, and graphical language used to define causal estimands and distinguish observation from intervention. Section 2.2 introduces the tail modelling tools used to extrapolate beyond observed data and to describe dependence among simultaneous extremes. The two frameworks are kept conceptually distinct here so that their roles are clear before they are combined in the methodological chapters.

2.1 Causal inference framework

2.1.1 Potential outcomes and causal estimands

Consider a population of units indexed by $i = 1, \dots, n$. Each unit can receive one of two treatments, encoded by a binary variable $D_i \in \{0, 1\}$, where $D_i = 1$ denotes receipt of the treatment and $D_i = 0$ denotes non-receipt. For each unit i and each treatment level $d \in \{0, 1\}$, the *potential outcome* $Y_i(d)$ represents the outcome that would be observed under treatment d. The observed outcome is

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0), \tag{2.1}$$

which connects the observed data to the underlying potential outcomes.

The central difficulty is that for any given unit, at most one potential outcome is observed: the one corresponding to the treatment actually received. The other potential outcome is *counterfactual*. This is the *fundamental problem of causal inference*. The framework therefore, treats $\{Y_i(0), Y_i(1)\}$ as a pair of random variables, only one element of which is revealed by the data, and seeks to recover features of the marginal distributions of $Y(0)$ and $Y(1)$, or contrasts between them, from observational or experimental data. Consequently, the joint distribution of $(Y(0), Y(1))$ is generally not identifiable (Rubin, 1974; Holland, 1986; Imbens and Rubin, 2015).

The potential outcomes notation is typically paired with two standard conditions. First, the *stable unit treatment value assumption* (SUTVA) rules out interference between units and hidden versions of treatment, so that $Y_i(d)$ depends only on unit i's own treatment status. Second, *consistency*, already implicit in (2.1), states that if $D_i = d$, then the observed outcome equals the corresponding potential outcome, $Y_i = Y_i(d)$.

Under these conditions, the observed outcome Y_i is determined by the treatment assignment and the two potential outcomes, and causal questions can be framed as comparisons between features of the distributions of $Y(1)$ and $Y(0)$. The choice of which features to compare—means, quantiles, or exceedance probabilities—defines the causal estimand and determines the statistical tools required for identification and estimation.

2.1.1.1 Treatment effects

The most common causal estimand is the *average treatment effect* (ATE), defined as

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]. \tag{2.2}$$

The ATE measures the expected difference in outcomes between the treated and control conditions, averaged over the entire population. A closely related quantity is the *average treatment effect on the treated* (ATT),

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid D = 1], \tag{2.3}$$

which restricts attention to the subpopulation that actually received the treatment. The ATE and ATT coincide when treatment is assigned randomly, but may differ under systematic selection into treatment.

While the ATE summarises the shift in location between the two potential outcome distributions, it provides no information about how the treatment affects other features of the distribution. The *quantile treatment effect* (QTE) at level $\tau \in (0,1)$ is defined as

$$\delta(\tau) = Q_{Y(1)}(\tau) - Q_{Y(0)}(\tau), \tag{2.4}$$

where $Q_{Y(d)}(\tau) = \inf\{y : F_{Y(d)}(y) \geq \tau\}$ is the τ-quantile of $Y(d)$. The QTE captures distributional heterogeneity: the treatment may have no effect on average but large effects at extreme quantiles, or vice versa. More generally, one may compare the entire distributions of $Y(1)$ and $Y(0)$ rather than a single summary. Distributional treatment effects refer to the collection of contrasts $\{F_{Y(1)}(y) - F_{Y(0)}(y) : y \in \mathbb{R}\}$ or, equivalently, to the set of QTEs $\{\delta(\tau) : \tau \in (0,1)\}$.

When attention turns to the tails, causal effects can be defined on extreme quantiles or exceedance probabilities. The *extreme quantile treatment effect* considers $\delta(\tau)$ as $\tau \to 1$ (or $\tau \to 0$). Alternatively, one may define causal effects on tail probabilities directly: for a high threshold y,

$$\mathbb{E}[\mathbb{1}\{Y(1) > y\}] - \mathbb{E}[\mathbb{1}\{Y(0) > y\}] = \Pr(Y(1) > y) - \Pr(Y(0) > y), \tag{2.5}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. As $y \to \infty$, these quantities probe the upper tail of the outcome distributions and are governed by the tail behaviour of $Y(1)$ and $Y(0)$ rather than by their central moments. The statistical challenge is that tail probabilities and extreme quantiles cannot be reliably estimated from moderate samples without additional structural assumptions on the tail—a topic addressed by extreme value theory in Section 2.2.

2.1.1.2 Causal attribution

A different family of causal questions concerns attribution rather than treatment comparison. In *extreme event attribution*, the target is the probability of a rare event under factual conditions (with the cause present) compared with its probability under counterfactual conditions (with the cause absent). Let F denote the factual world and CF the counterfactual world. For a high threshold v, define the event $E = \{Y > v\}$. The *risk ratio* is

$$\mathrm{RR} = \frac{\Pr(Y > v \mid F)}{\Pr(Y > v \mid CF)} = \frac{p_1}{p_0}, \tag{2.6}$$

where $p_1 = \Pr(Y > v \mid F)$ and $p_0 = \Pr(Y > v \mid CF)$. A risk ratio greater than one indicates that the cause increased the probability of the event; the magnitude quantifies the extent of the increase.

A complementary measure is the *probability of necessary causation* (PN), defined as

$$\mathrm{PN} = \Pr(Y_0 \leq v \mid Y_1 > v), \tag{2.7}$$

i.e., the probability that the event would not have occurred absent the cause, given that it did occur. Under the monotonicity assumption that $Y_1 \geq Y_0$ almost surely, this reduces to

$$\mathrm{PN} = \max\left(1 - \frac{p_0}{p_1},\, 0\right). \tag{2.8}$$

Without monotonicity, the right-hand side serves only as a lower bound. Related quantities include the *probability of sufficient causation*, $\mathrm{PS} = \Pr(Y_1 > v \mid Y_0 \leq v)$, and the *probability of necessary and sufficient causation*, $\mathrm{PNS} = \Pr(Y_1 > v, Y_0 \leq v)$, whose identification likewise depends on the underlying counterfactual assumptions.

Both treatment effects and attribution metrics compare outcomes under alternative states of the world, but the relevant contrasts differ. Treatment effect estimation contrasts treatment levels, often under observational identification assumptions such as unconfoundedness. Attribution instead contrasts factual and counterfactual generating mechanisms, and therefore often relies on physical or simulation models to construct the counterfactual world.

2.1.2 Identification strategies

Returning to the potential outcomes notation of Section 2.1.1, this section reviews the assumptions and estimation strategies that connect observed data to causal quantities.

2.1.2.1 Identification assumptions

For observational data, identification of the ATE and related estimands typically rests on consistency, SUTVA, unconfoundedness, and overlap, with unconfoundedness and overlap being the distinctively observational assumptions.

The *unconfoundedness* assumption, also called conditional independence or ignorability, states that

$$\big(Y(0), Y(1)\big) \perp\!\!\!\perp D \mid \boldsymbol{X}, \tag{2.9}$$

where $\boldsymbol{X}$ is a vector of observed pre-treatment covariates. Conditional on $\boldsymbol{X}$, treatment assignment is as good as random. This assumption is untestable from the data alone and must be justified substantively.

The *overlap* or *positivity* assumption requires that

$$0 < \Pr(D = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1 \quad \text{for all } \boldsymbol{x} \text{ in the support of } \boldsymbol{X}. \tag{2.10}$$

This ensures that every covariate profile has positive probability of receiving either treatment, so that both potential outcomes are represented in principle. Violations of overlap force causal comparisons to rely on extrapolation.

Together with consistency and SUTVA, unconfoundedness and overlap permit identification of the ATE:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}[Y \mid D = 1, \boldsymbol{X}] - \mathbb{E}[Y \mid D = 0, \boldsymbol{X}]\big]. \tag{2.11}$$

The right-hand side depends only on the observed data distribution, since under unconfoundedness $\mathbb{E}[Y \mid D = d, \boldsymbol{X}] = \mathbb{E}[Y(d) \mid \boldsymbol{X}]$, and iterating expectations over $\boldsymbol{X}$ recovers the marginal causal contrast. This expression of the ATE in terms of observable conditional means is the canonical identification result and underlies the estimation strategies developed below.

2.1.2.2 Inverse probability weighting

The *propensity score* is defined as $\pi(\boldsymbol{X}) = \Pr(D = 1 \mid \boldsymbol{X})$. Under unconfoundedness, conditioning on the propensity score suffices to remove confounding i.e., $\big(Y(0), Y(1)\big) \perp\!\!\!\perp D \mid \pi(\boldsymbol{X})$ (Rosenbaum and Rubin, 1983a). *Inverse probability weighting* (IPW) exploits this property to construct an unbiased estimator of $\mathbb{E}[Y(d)]$. Under unconfoundedness and overlap, we have that, for $d \in \{0, 1\}$,

$$\mathbb{E}[Y(d)] = \mathbb{E}\left[\frac{\mathbb{1}(D = d)\, Y(d)}{\Pr(D = d \mid \boldsymbol{X})}\right]. \tag{2.12}$$

The sample analogue replaces the true propensity score with an estimate $\hat{\pi}(\boldsymbol{X})$, yielding

$$\hat{\mathbb{E}}[Y(1)] = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{\hat{\pi}(\boldsymbol{X}_i)}, \qquad \hat{\mathbb{E}}[Y(0)] = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - D_i) Y_i}{1 - \hat{\pi}(\boldsymbol{X}_i)}. \tag{2.13}$$

IPW estimators are intuitive and widely used but can suffer from high variance when estimated propensity scores are close to zero or one (Kang and Schafer, 2007).

2.1.2.3 Estimating equations

A more general estimation strategy expresses the causal estimand as the solution to a population-level moment condition. For a scalar parameter θ and an estimating function $g(W, \theta, \zeta)$, where W denotes the observed data and ζ collects nuisance parameters, the *estimating equation* is

$$\mathbb{E}[g(W, \theta, \zeta)] = 0. \tag{2.14}$$

The estimator $\hat{\theta}$ is obtained by solving the empirical analogue

$$\frac{1}{n}\sum_{i=1}^{n} g(W_i, \hat{\theta}, \hat{\zeta}) = 0. \tag{2.15}$$

This framework encompasses IPW (where g is a weighted indicator function), outcome regression, and hybrid approaches (Tsiatis, 2006). Its generality is important for the extension to extreme quantile estimation developed in Chapter 5, where the estimating function is tailored to aggregate information across quantile levels and incorporate tail calibration.

2.1.2.4 Other standard strategies

Two further strategies are standard in causal inference but play a more peripheral role in this thesis. *Doubly robust* estimators combine an outcome model with a propensity score model and remain consistent if either component is correctly specified (Robins et al., 1994). *Instrumental variable* methods instead address settings where unconfoundedness is not credible by exploiting variables that shift treatment assignment without directly affecting the outcome (Angrist et al., 1996). Both belong to the standard causal toolkit, but neither is central to the methodological developments in the later chapters.

2.1.3 Structural causal models and their graphical representation

The potential outcomes framework provides a language for defining causal effects but does not, by itself, specify the mechanisms by which causes produce outcomes. *Structural causal models* (SCMs) offer a complementary perspective that makes the data-generating process explicit. A directed acyclic graph then provides a graphical representation of the structural relationships encoded by the model (Pearl, 2009b; Spirtes et al., 2000; Peters et al., 2017).

An SCM consists of three components: (i) a set of endogenous variables $\boldsymbol{V} = \{V_1, \ldots, V_p\}$ whose values are determined within the model; (ii) a set of exogenous variables $\boldsymbol{U} = \{U_1, \ldots, U_p\}$ representing sources of randomness external to the model; and (iii) a set of structural functions $\{f_j\}_{j=1}^{p}$ such that

$$V_j = f_j(\mathrm{pa}(V_j), U_j), \qquad j = 1, \ldots, p, \tag{2.16}$$

where $\mathrm{pa}(V_j) \subseteq \boldsymbol{V} \backslash \{V_j\}$ denotes the *parents* of V_j in the model and U_j is an exogenous noise term. Each structural function encodes a direct causal mechanism: V_j is determined by its parents and its own exogenous noise, holding fixed all other variables. For expositional clarity, we present the Markovian case in which each U_j enters only V_j and the U_j are mutually independent; this assumption is relaxed in Chapter 6, where latent common shocks are permitted.

The parent relationships in an SCM induce a *directed acyclic graph* (DAG) $\mathcal{G} = (\boldsymbol{V}, E)$, whose nodes are the endogenous variables $\boldsymbol{V}$ and whose edge set E contains a directed edge $V_i \to V_j$ if and only if $V_i \in \text{pa}(V_j)$. The graph is acyclic, meaning no variable is its own ancestor. In this sense, the DAG is not a separate causal object but a compact representation of the structural assumptions encoded by the SCM. A path is a sequence of adjacent nodes in the underlying graph, irrespective of the direction of the arrows along that sequence. The graph also provides a basis for reading off conditional independence relations via the notion of *d-separation*: a path between two nodes is *blocked* by a set of nodes S if the path contains a chain $V_i \to V_m \to V_j$ or a fork $V_i \leftarrow V_m \to V_j$ with $V_m \in S$, or a collider $V_i \to V_m \leftarrow V_j$ with neither V_m nor any descendant of V_m in S. Two nodes are *d-separated* by S if every path between them is blocked; under the Markov assumption, d-separation implies conditional independence. Figure 2.1 illustrates the three triple configurations and the corresponding blocking rules.

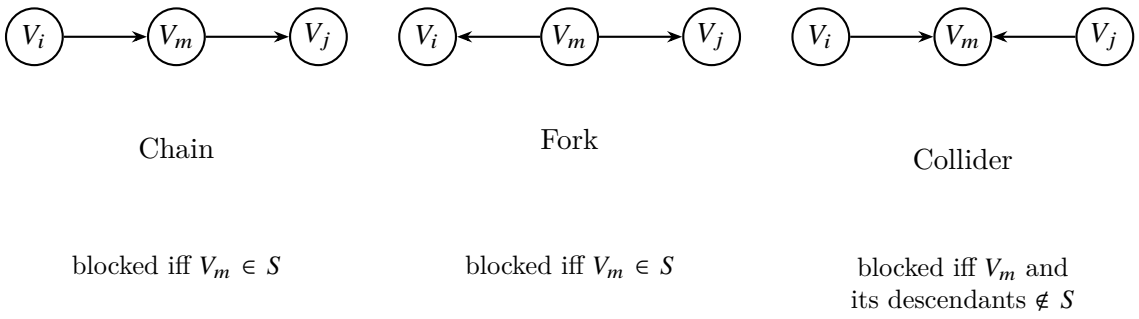

Figure 2.1: The three triple configurations underlying d-separation (Pearl, 2009b). A chain or fork transmits dependence between V_i and V_j unless the middle node V_m is included in the conditioning set S. A collider behaves in the opposite way, with the path blocked by default and opened only when V_m or one of its descendants is included in S.

The graphical framework also clarifies the concept of *confounding*. A variable is a common cause of X and Y if it is an ancestor of both variables in the DAG. A variable C is a confounder of the effect of X on Y if C is a common cause of both X and Y. In the DAG, this creates a *backdoor path* $X \leftarrow C \to Y$ that generates spurious association

between X and Y unrelated to the causal effect of interest. The *backdoor criterion* states that if there exists a set of variables S that blocks all backdoor paths from X to Y without containing any descendant of X, then conditioning on S is sufficient to identify the causal effect of X on Y.

The SCM framework supports the *do-operator*, which represents a hypothetical intervention. The expression $\Pr(Y = y \mid \mathrm{do}(X = x))$ denotes the probability that $Y = y$ under an intervention that sets X to x, removing the influence of all parents of X. This contrasts with the conditional probability $\Pr(Y = y \mid X = x)$, which reflects passive observation and may be distorted by confounding. More generally, *do-calculus* provides graphical rules for expressing interventional distributions in terms of observational quantities when identification is possible.

2.2 Extreme value theory

This section reviews the probabilistic foundations of extreme value theory, covering the characterisation of tail behaviour, the two classical paradigms for modelling univariate extremes, and the extension to multivariate settings with extremal dependence (Coles, 2001; Beirlant et al., 2004; De Haan and Ferreira, 2007).

2.2.1 Tail behaviour and regular variation

Many problems in statistical modelling concern rare and extreme events rather than typical outcomes, including large financial losses, daily temperature maxima, and structural loads at design return periods. Reliable inference about such events is governed by the tail of the distribution, where central-limit-style approximations no longer apply. Extreme value theory supplies both a language for tail behaviour and the limit theorems that govern extremes.

A central distinction in distributional theory is between *light-tailed* and *heavy-tailed* distributions. Loosely, distributions with polynomially decaying tails form a canonical class of heavy-tailed models, so that extreme values remain relatively likely even far from the centre. By contrast, many light-tailed distributions decay exponentially or faster, though the precise classification is more naturally expressed through domain-of-attraction and regular-variation arguments.

A precise and widely used formulation is the theory of *regular variation*. A distribution function F on $[0, \infty)$ with tail $\bar{F}(y) = 1 - F(y)$ is *regularly varying* with index $\alpha > 0$, written $\bar{F} \in \mathrm{RV}_{-\alpha}$, if

$$\lim_{t \to \infty} \frac{\bar{F}(ty)}{\bar{F}(t)} = y^{-\alpha}, \qquad \text{for all } y > 0. \tag{2.17}$$

Equivalently, $\bar{F}(y)$ has Pareto-type decay, in the sense that $\bar{F}(y) \approx L(y)y^{-\alpha}$ for some slowly varying function L satisfying $L(ty)/L(t) \to 1$ as $t \to \infty$. Regularly varying distributions are heavy-tailed and include the Pareto, Fréchet, Student-t, and Cauchy families (Resnick, 2007).

Regular variation describes how the parent distribution's tail decays. The natural next question is what this implies for the largest observed values. If $Y_1, \ldots, Y_n$ are iid with distribution F, the problem is to characterise how the maximum $M_n = \max(Y_1, \ldots, Y_n)$ behaves as $n \to \infty$. The answer is given by the theorem of Fisher, Tippett, and Gnedenko, the cornerstone of univariate extreme value theory.

2.2.2 Univariate extremes

Extreme value theory provides two principal routes to statistical modelling of tail behaviour, namely the block maxima approach and the peaks-over-threshold approach.

2.2.2.1 Block maxima and extreme value limits

The block maxima approach divides the data into blocks and models the maximum within each block. Its theoretical foundation is the theorem of Fisher, Tippett, and Gnedenko, which characterises the possible non-degenerate limits of normalised maxima. If $M_n = \max(Y_1, \ldots, Y_n)$ and there exist normalising sequences $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\lim_{n \to \infty} \Pr\left(\frac{M_n - b_n}{a_n} \leq y \right) = G(y), \tag{2.18}$$

where G is non-degenerate, then F is said to belong to the *max-domain of attraction* (MDA) of G, and G must coincide, up to location and scale, with one of the following three parametric types.

- The *Fréchet* case ($\gamma > 0$) has heavy, polynomially decaying tails.

- The *Gumbel* case ($\gamma = 0$) has light, exponentially decaying tails and includes the normal, exponential, and lognormal.
- The *Weibull* case ($\gamma < 0$) has bounded tails with a finite upper endpoint.

The parameter γ, called the *extreme value index* or *shape parameter*, controls the rate of tail decay and the rate at which extreme quantiles diverge; it is the single most important quantity for extrapolation beyond the observed data. The three types are unified by the *generalised extreme value* (GEV) distribution, which has location, scale, and shape parameters.

The connection back to regular variation is direct. A distribution F with $\bar{F} \in \mathrm{RV}_{-\alpha}$ for $\alpha > 0$ lies in the Fréchet MDA with shape $\gamma = 1/\alpha$. Polynomial tail decay of the parent thus determines both the existence of a non-degenerate limit law for M_n and the sign and magnitude of the limiting shape parameter.

In this thesis the block maxima formulation is included only as standard background; the later chapters rely more directly on threshold exceedances and tail extrapolation.

2.2.2.2 Threshold exceedances and the generalised Pareto distribution

An alternative approach uses exceedances over a high threshold u. For a random variable Y with distribution F, the *conditional excess distribution* is

$$F_u(y) = \Pr(Y - u \le y \mid Y > u) = \frac{F(u+y) - F(u)}{1 - F(u)}, \qquad y > 0. \tag{2.19}$$

The theorem of Balkema, de Haan, and Pickands states that for a large class of distributions F in the MDA of G, the excess distribution converges to the *generalised Pareto distribution* (GPD) as $u \to y^* = \sup\{y : F(y) < 1\}$ (Balkema and de Haan, 1974; Pickands, 1975), as shown by

$$\sup_{0<y<y^*-u} |F_u(y) - H(y; \sigma_u, \gamma)| \to 0 \quad \text{as } u \to y^*, \tag{2.20}$$

where

$$H(y;\sigma_u,\gamma)=\begin{cases}1-\left(1+\dfrac{\gamma y}{\sigma_u}\right)^{-1/\gamma}, & \gamma\neq 0,\quad y>0,\quad 1+\gamma y/\sigma_u>0,\\ 1-\exp\left(-\dfrac{y}{\sigma_u}\right), & \gamma=0,\quad y>0.\end{cases} \tag{2.21}$$

The GPD is parameterised by a scale $\sigma_u > 0$ (which depends on the threshold) and the shape γ (which is the same γ as in the GEV). For $\gamma > 0$, the GPD has a heavy, Pareto-type tail; for $\gamma = 0$, an exponential tail; and for $\gamma < 0$, a bounded upper endpoint at $-\sigma_u/\gamma$.

The peaks-over-threshold approach is often preferred in practice because it uses data more efficiently than block maxima, since all observations exceeding the threshold contribute to the fit rather than only the single largest value per block. The threshold u must be chosen high enough for the GPD approximation to be accurate but low enough to retain sufficient exceedances for stable estimation. Diagnostics such as mean excess plots and threshold stability plots are used to guide this choice.

2.2.2.3 Extreme quantiles and return levels

One of the central applications of EVT is the estimation of quantiles at levels τ beyond the range of the data. Let $\hat{F}$ denote the empirical distribution and let u be a high threshold such that $n_u = \sum_{i=1}^{n} \mathbb{1}(Y_i > u)$ exceedances are observed. For $\tau > \hat{F}(u)$, the extreme quantile is estimated by inverting the following semi-parametric model, which combines the empirical distribution below u with the GPD above u.

$$\hat{Q}(\tau) = u + \frac{\hat{\sigma}_u}{\hat{\gamma}}\left[\left(\frac{n(1-\tau)}{n_u}\right)^{\hat{\gamma}} - 1\right], \tag{2.22}$$

where $\hat{\gamma}$ and $\hat{\sigma}_u$ are GPD parameter estimates. This estimator extrapolates beyond the largest observation using the tail shape, with heavy tails ($\hat{\gamma} > 0$) producing faster growth and light tails ($\hat{\gamma} = 0$) yielding exponential extrapolation. The quality of the extrapolation depends critically on the accuracy of $\hat{\gamma}$, which is typically the most difficult parameter to estimate.

A closely related quantity is the *return level*. For a return period of m observations, the return level z_m is defined by $\Pr(Y > z_m) = 1/m$, equivalently $z_m = Q(1 - 1/m)$. Return levels translate tail probabilities into the scale of the original outcome and are often more interpretable in applications. In the present thesis, extreme quantiles and return levels are especially relevant because later chapters focus on tail probabilities, threshold exceedances, and causal effects defined in the extreme regime rather than on central features of the outcome distribution.

2.2.3 Multivariate extremes and extremal dependence

Univariate EVT characterises the marginal tail behaviour of individual variables. In many applications, however, the joint behaviour of multiple variables in the tail is of primary interest, as in a heatwave coinciding with a drought, co-movements of financial losses across markets, or simultaneous flooding at multiple gauges. Joint tail behaviour is shaped by both marginal heaviness and *extremal dependence*, and the two must be modelled together. This section summarises the standard framework for doing so.

2.2.3.1 Multivariate regular variation and marginal standardisation

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ be a d-dimensional random vector with distribution F. In multivariate extremes, the first step is usually to standardise the margins to a common tail scale. Let T_j be a monotone transformation such that $T_j(Y_j)$ follows a standard heavy-tailed distribution, for example unit Pareto or unit Fréchet margins, and write $\boldsymbol{Z} = (T_1(Y_1), \ldots, T_d(Y_d))$. This standardisation separates marginal tail behaviour from cross-sectional extremal dependence and makes dependence comparisons meaningful across coordinates.

A common framework for studying the tail of $\boldsymbol{Z}$ is *multivariate regular variation*. In its simplest form, this means that there exists a non-null measure ν such that for suitable sets A bounded away from the origin,

$$t \Pr(\boldsymbol{Z}/t \in A) \to \nu(A), \qquad t \to \infty. \tag{2.23}$$

The limit measure ν describes the relative frequency of different extremal directions and magnitudes. This framework underlies both threshold exceedance modelling and tail dependence analysis because it provides the mathematical basis for discussing whether extremes tend to occur jointly, whether mass concentrates on particular faces of the state space, and how extremal dependence persists after marginal standardisation (De Haan and Ferreira, 2007; Rootzén et al., 2018).

A complementary block maxima perspective considers componentwise maxima and leads to max-stable limit distributions, which play the role of multivariate extreme value limits. This viewpoint is part of standard multivariate EVT, but it is not the main route used in this thesis. The later chapters rely more directly on threshold exceedances, marginal standardisation, and extremal dependence summaries, so max-stability is mentioned here only to situate regular variation within the broader EVT literature.

2.2.3.2 Extremal dependence coefficients

A widely used summary of bivariate extremal dependence is the *extreme dependence coefficient*, defined for a pair (Y_1, Y_2) as

$$\chi = \lim_{\tau \to 1} \Pr(Y_2 > Q_{Y_2}(\tau) \mid Y_1 > Q_{Y_1}(\tau)), \tag{2.24}$$

provided the limit exists. The coefficient χ lies in $[0, 1]$ and measures the limiting probability that one variable is extreme given that the other is extreme. Values closer to one indicate stronger extremal dependence. A value of $\chi = 0$ indicates *asymptotic independence*, while $\chi = 1$ indicates perfect *asymptotic dependence*. A complementary measure is (Coles et al., 1999)

$$\bar{\chi} = \lim_{\tau \to 1} \frac{2 \log \Pr(Y_1 > Q_{Y_1}(\tau))}{\log \Pr(Y_1 > Q_{Y_1}(\tau), Y_2 > Q_{Y_2}(\tau))} - 1, \tag{2.25}$$

which captures residual dependence even when $\chi = 0$.

The distinction between *asymptotic dependence* and *asymptotic independence* has important consequences for tail modelling. In terms of the coefficient χ, asymptotic independence corresponds to $\chi = 0$, while asymptotic dependence corresponds to $\chi > 0$; the special case $\chi = 1$ represents perfect asymptotic dependence. Under asymptotic depend-

ence, the joint tail probability $\Pr(Y_1 > y, Y_2 > y)$ decays at the same rate as the marginal $\Pr(Y_1 > y)$, so extreme events in one variable remain informative about extremes in the other. Under asymptotic independence, the joint tail decays faster than the marginal, meaning that simultaneous extremes become increasingly rare relative to individual extremes.

Chapter 3
Literature review

This chapter surveys the existing literature on extreme event attribution, causal inference on extreme outcomes, and causal discovery in heavy-tailed settings. The three problems are treated not as independent topics but as successive inferential dependencies, and the review is organised to expose the gaps that motivate the contributions of the thesis.

3.1 The divide between extremes and causality

The literatures on extremes and on causality have largely developed apart. Extreme event attribution, extreme quantile treatment effect estimation, and causal discovery in multivariate extremes all sit at their intersection, yet they are often treated as separate methodological problems rather than as connected responses to the same inferential divide. In this thesis, they are instead understood as linked tasks. Attribution requires estimating causal effects on extreme outcomes. Estimating such effects, in turn, requires either knowing the causal structure that generated the data or adjusting for it correctly. Recovering that structure from extreme observations then requires methods adapted to the geometry of extremal dependence. Each task therefore creates conditions under which the previous one can be carried out credibly.

This chapter reviews that pipeline from endpoint to foundation. Section 3.2 examines the extreme event attribution literature, highlighting both the dependence sensitivity of current methods and their reliance on climate model output rather than observational data. Section 3.3 turns to extreme quantile treatment effect estimation, showing why standard quantile methods become unstable in the tail and how recent work has begun to combine extreme value theory with causal inference under a known causal structure.

Section 3.4 then reviews causal discovery methods, both classical and extremal, and identifies the remaining gap in recovering directed structure from heavy-tailed data. Section 3.5 brings these strands together and argues that their apparent separation has obscured a single integrated inferential problem.

Detailed presentations of the methods developed in this thesis are deferred to the later substantive chapters. The purpose here is instead to establish the chapter-level narrative of the thesis and to show why progress on attribution depends on progress on estimation, and why progress on estimation depends on discovery.

3.2 Attribution as retrospective causal reasoning

3.2.1 The attribution landscape

Extreme event attribution is a retrospective form of causal reasoning. It asks, after an extreme event has occurred, whether anthropogenic climate change altered its probability or intensity. As a scientific discipline, extreme event attribution (EEA) emerged in the early 2000s. Allen (2003) framed the problem in terms of legal and scientific liability, arguing that the same physical models used to project future climate could be employed to assess the causal contribution of greenhouse gas emissions to observed events. Stott et al. (2004) provided the first formal attribution of a specific event, estimating that human influence had at least doubled the probability of the 2003 European summer heatwave. The methodological foundations were consolidated through the IPCC detection and attribution framework (Bindoff et al., 2013; Hegerl et al., 2010), which established good practice guidelines for formal attribution studies. The National Academies of Sciences, Engineering, and Medicine (2016) report subsequently established attribution as a mature scientific enterprise, providing confidence frameworks and evaluation standards. Clarke et al. (2022) provided a comprehensive review of attribution across hazard types, documenting the rapid expansion of the field from temperature and precipitation events to tropical cyclones, wildfires, and compound events.

Philip et al. (2020) formalised the operational protocol used by the World Weather Attribution initiative, establishing standardised procedures for rapid attribution of extreme events within days of their occurrence. This protocol has been applied to hundreds of events worldwide, demonstrating the feasibility of attribution on policy-relevant timescales.

The *storyline* or *circulation-based* approach conditions on the observed large-scale atmospheric circulation and asks how thermodynamic changes, primarily increased moisture capacity and temperature shifts, have altered the intensity or conditional likelihood of the event within that specific weather pattern (Shepherd, 2016; Shepherd et al., 2018; Trenberth et al., 2015). This approach avoids the difficulty of simulating the counterfactual circulation and instead attributes changes to the component of the event most directly linked to global warming, namely thermodynamic amplification. Its strength lies in the physical interpretability of the results for a specific event, but by conditioning on circulation it does not, on its own, quantify changes in the event's overall probability, since the circulation pattern itself may also be influenced by climate change. Sippel et al. (2020) showed that the anthropogenic climate change signal is now detectable from a single day of global weather, highlighting the strength of the underlying thermodynamic signal. Jézéquel et al. (2018) further illustrated how thermodynamic and dynamic contributions can be separated in the analysis of extreme heat events.

The *probabilistic event-attribution* approach, which is the framework used in Chapter 4, estimates the probability of a defined event under two scenarios: a factual world with anthropogenic forcing (p_1) and a counterfactual world without it (p_0). It computes metrics such as the risk ratio p_1/p_0, the fraction of attributable risk $1 - p_0/p_1$, and the probability of necessary causation (Hannart et al., 2016; Pearl, 2009a). Estimating the small probabilities p_1 and p_0 places heavy demands on both climate models and statistical tail extrapolation.

Both the storyline and probabilistic event-attribution approaches are typically implemented using climate model simulations. The storyline approach requires simulations that resolve the relevant weather patterns under altered thermodynamic conditions, whereas probabilistic event attribution generally requires ensembles under factual and counterfactual forcings. These analyses therefore depend on the availability and reliability of climate

model output, which may be limited in regional or data-sparse settings and for compound events involving multiple hazards. This reliance on simulated counterfactuals, rather than on observational data, is a structural limitation of current attribution practice. In parallel, a growing body of work has explored attribution using observational and reanalysis data, aiming to reduce reliance on model-based counterfactuals. For example, Fischer and Knutti (2015) showed that a large fraction of observed heavy precipitation and heat extremes is attributable to human-induced warming, while Risser and Wehner (2017) quantified the human contribution to extreme precipitation during Hurricane Harvey using observational records and statistical methods. Otto et al. (2012) further demonstrated how thermodynamic and dynamic contributions can be jointly assessed in specific events. While these studies strengthen the empirical basis for attribution, they also highlight the statistical challenges of estimating tail probabilities from finite records.

A distinct challenge, increasingly recognised in the literature, is that many impactful extreme events are compound and arise from the co-occurrence of multiple hazards, such as concurrent heat and drought or co-occurring precipitation at multiple locations. Zscheischler et al. (2020) provided a systematic typology of compound weather and climate events, highlighting that their risk is governed by the joint tail behaviour of the contributing variables. Cattiaux and Ribes (2018) further argued that the definition of extreme events itself must account for the multivariate and spatial structure of the climate system. Compound event attribution places even greater demands on the dependence model, since the joint tail probability depends on how multiple variables co-exceed their respective thresholds.

Two unresolved issues emerge from this discussion and motivate the remainder of the section. The centrality of joint tail behaviour for compound events raises the sensitivity of attribution conclusions to the choice of extremal dependence model. The reliance on simulated rather than observational counterfactuals points to the limited role that observational causal inference currently plays in attribution practice.

3.2.2 Probabilistic attribution with extreme value models

Within the probabilistic approach, extreme value theory provides the natural statistical framework for estimating tail probabilities. The classical results underpinning this framework are well established. As we saw in Chapter 2, the generalised extreme value (GEV) distribution arises as the limiting law of block maxima (Coles, 2001), and the generalised Pareto distribution (GPD) approximates the distribution of exceedances above a high threshold (Balkema and de Haan, 1974; Pickands, 1975). These results provide principled extrapolation beyond the observed data range, which is essential when the event of interest is more extreme than any observation on record.

Univariate EVT has been widely applied in attribution. Van Der Wiel et al. (2017) fitted the GEV distribution to annual maxima of three-day precipitation averages for rapid attribution of the 2016 Louisiana floods. Nanditha et al. (2024) used an extended GPD to attribute daily precipitation extremes across the United States to anthropogenic emissions, avoiding arbitrary threshold selection. Gonzalez et al. (2025) introduced a framework for modelling non-stationarity in precipitation records and quantifying changes in record-breaking events over time. These methods work well for single-site attribution but treat each location or variable independently, ignoring the dependence between sites. In practice, extreme events are often spatially extended, and a precipitation event that is exceptional at one gauge is typically exceptional at neighbouring gauges as well. The probability of a spatially compound event therefore depends critically on how extremal dependence is modelled.

3.2.3 Dependence sensitivity and the causal gap

Kiriliouk and Naveau (2020) were among the first to incorporate multivariate extreme value models into the attribution framework. They used the multivariate generalised Pareto distribution (mGPD) to model joint tail behaviour when at least one component of a multivariate vector exceeds a high threshold, combining it with Pearl's counterfactual causation theory. Their work demonstrated that multivariate EVT can substantially change attribution conclusions relative to marginal-only analyses. However, the mGPD assumes asymptotic dependence, meaning that extremal dependence persists even at the most extreme levels. That assumption is violated in many environmental applications

where dependence weakens as events become more severe. Alternative models, including asymptotic independence formulations and subasymptotic models (Huser and Wadsworth, 2019; Castro-Camilo and Huser, 2020), can yield markedly different estimates of p_0 and p_1 for the same event, even when the marginal tail specifications are identical.

Naveau et al. (2020a) surveyed the dominant statistical approaches to attribution and noted that most existing methods treat marginal distributions and dependence structure as separate modelling decisions, rarely interrogating how sensitive the resulting attribution statements are to the latter. The dependence structure determines whether multiple locations or variables co-exceed their thresholds simultaneously, and therefore what counts as the same event across the factual and counterfactual worlds.

This dependence sensitivity points to a deeper issue. Attribution is ultimately a causal inference problem, but current probabilistic methods still rely heavily on model-generated counterfactuals rather than on observational identification. A broader question then arises, namely whether causal effects on extreme outcomes can be estimated directly from observational data. This shift in viewpoint moves the focus from simulating counterfactual worlds to identifying causal effects from data, but also introduces new statistical challenges, as the outcomes of interest lie in the tail of the distribution.

3.3 Estimation as prospective inference on extreme outcomes

3.3.1 From attribution to the need for observational causal inference

Where attribution asks how an observed extreme event came about, estimation asks how changing a treatment or intervention would alter extreme outcomes before they are observed. It is therefore a prospective causal problem, and in practice it often must be addressed from observational data. Both dominant attribution approaches, namely storyline and probabilistic attribution, depend on climate model output to construct counterfactual worlds. The storyline approach requires simulations that resolve specific weather patterns under altered thermodynamic conditions, while the probabilistic approach requires large

ensembles under factual and counterfactual forcings. In many practical settings, however, such output is unavailable or unreliable. Regional models may lack the resolution needed for local attribution, observational records may be short, and compound events can exceed the capacity of even state-of-the-art climate models.

These limitations motivate a broader reformulation of the problem. Rather than asking whether a particular climate model configuration attributes an event to anthropogenic forcing, one can ask whether observational data alone provide evidence that a treatment variable, such as time period, forcing level, or policy regime, causally shifts the extremal behaviour of an outcome. Framed this way, attribution becomes a problem of causal inference from observational data, with the estimand defined on extreme quantiles or exceedance probabilities rather than on averages.

The challenges are immediate. Extreme outcomes are by definition rare, so the effective sample size in the tail is small. Standard overlap conditions, which ensure adequate covariate support across treatment groups, may hold comfortably for median outcomes but fail for extreme quantiles. The very strata where tail behaviour must be characterised may contain no observations at all. Extrapolation beyond the observed data range becomes necessary, and the form of that extrapolation is governed by the tail of the distribution, which is precisely the quantity that is hardest to estimate.

3.3.2 Causal inference and quantile treatment effects

The potential outcomes framework, originating with Neyman (1990, originally 1923) (The 1990 English translation follows the 1923 Dutch original.) and developed by Rubin (1974) into its modern form, provides the standard formalism for causal inference from observational data. The key identifying assumptions of unconfoundedness, overlap, and consistency were formalised in the context of treatment effect estimation by Rosenbaum and Rubin (1983b) through the propensity score, and subsequently elaborated in the comprehensive treatments of Angrist and Pischke (2009) and Hernán and Robins (2020). For a binary treatment $D \in \{0, 1\}$, the potential outcomes $Y(0)$ and $Y(1)$ represent the outcome that would be observed under each treatment state. The average treatment effect

$\mathbb{E}[Y(1) - Y(0)]$ captures the mean difference but provides no information about how the treatment affects different parts of the outcome distribution. When the scientific interest lies in the tails, as it does for extreme events, the average treatment effect is the wrong estimand.

Quantile treatment effects (QTEs) provide a more informative alternative. Doksum (1974) and Lehmann and D'Abrera (1975) introduced the concept of comparing quantiles of the treatment and control distributions as a measure of distributional treatment impact. Koenker and Bassett (1978) established the theory of regression quantiles, providing the computational and statistical machinery that underpins modern quantile-based methods; see Koenker (2005) for a comprehensive treatment. Chamberlain (1994) linked quantile regression to structural models, and Firpo (2007) developed efficient semiparametric estimators for QTEs under unconfoundedness, using inverse probability weighting to construct a reweighted quantile loss function. Ma et al. (2023) extended Firpo's approach by linking marginal and conditional quantiles through quantile regression, improving precision. Donald and Hsu (2014) provided estimation and inference procedures for distribution and quantile functions in treatment effect models, and Frölich and Melly (2013) developed methods for unconditional QTE estimation under endogeneity. These methods perform well for interior quantile levels where the data provide adequate support across treatment groups. The limitation of average effects for capturing distributional impacts has been emphasised by Bitler et al. (2006), who showed that welfare reform experiments had heterogeneous effects across the outcome distribution that were entirely invisible to the average treatment effect.

The causal identification of QTEs relies on the same assumptions as average treatment effect estimation, namely unconfoundedness, overlap, and consistency, but their practical implications are different at extreme levels. Unconfoundedness requires that all confounders are observed, while overlap requires that every covariate stratum has a positive probability of receiving each treatment. At the median, these conditions are often plausible. At the 99th percentile, the set of relevant confounders may differ from those important at the centre, and the overlap condition may fail because certain covariate patterns that are common in one treatment group may produce outcomes so extreme that they are never observed in the other.

3.3.3 Why quantile methods fail in the tail

Zhang (2018a) formalised the difficulty of extreme quantile estimation by distinguishing three asymptotic regimes for quantile levels $\tau_n \to 0$. In the *intermediate* regime ($n\tau_n \to \infty$), the effective sample size in the tail is still growing, and Gaussian limits for suitably normalised estimators can be recovered. In the *moderately extreme* regime ($n\tau_n \to d > 0$), empirical quantiles become unstable and standard bootstrap procedures fail. In the *extreme* regime ($n\tau_n \to 0$), few or no observations fall in the tail region, and pointwise quantile inversion is fundamentally ill-posed. The practical consequence is that no amount of data will resolve the most extreme quantiles through empirical methods alone.

Chernozhukov (2005) established the foundational theory for extremal quantile regression, showing that convergence rates at extreme quantile levels degrade to logarithmic order under standard regularity conditions. This is a fundamental bottleneck. Chernozhukov and Fernández-Val (2011) developed inference procedures for extremal conditional quantile models, and Chernozhukov et al. (2018) provided an overview of the field. These results demonstrate that extremal quantile regression is theoretically possible but that the rate of convergence is inherently slower than at interior levels, motivating the use of explicit tail models to improve efficiency. Outside the causal setting, Gardes and Girard (2010) and Daouia et al. (2013) developed kernel-based and conditional methods for estimating extreme quantiles from heavy-tailed distributions, establishing nonparametric alternatives that avoid parametric tail specifications. Portnoy (2003) developed regression quantile methods for censored data, relevant in extreme settings where observations may be truncated by measurement limits or detection thresholds.

3.3.4 EVT-based tail extrapolation meets causal inference

Extreme value theory provides the natural framework for extrapolation beyond the observed data range. The foundational results are due to Balkema and de Haan (1974) and Pickands (1975), who established that the distribution of exceedances above a sufficiently high threshold converges to a generalised Pareto form, with comprehensive textbook treatments in De Haan and Ferreira (2007) and Coles (2001). For heavy-tailed distributions, the

tail quantile function follows a power-law representation governed by the extreme value index, typically estimated using classical procedures such as the Hill estimator (Hill, 1975). These tools enable estimation of quantiles beyond the range of observed data, provided the tail behaviour is correctly characterised.

Integrating EVT with causal inference for treatment effects is a more recent development. Deuber et al. (2024) proposed causal Hill estimators for the extreme value indices of potential outcome distributions, using inverse probability weighting to construct weighted tail estimators, and extrapolating intermediate quantile estimates to extreme levels via a power-law tail approximation. This approach provides a principled two-step procedure but restricts attention to heavy-tailed distributions in the Fréchet domain of attraction and does not integrate causal adjustment with tail calibration within a single estimating equation. Bhuyan et al. (2023) proposed a copula-based approach that reconstructs counterfactual outcome distributions using a bulk-tail decomposition with a data-driven transition threshold, demonstrating improved performance over standard quantile regression, but the framework relies on parametric copula specifications and does not provide a general treatment of tail extrapolation uncertainty. Cheng and Li (2025) introduced the inverse estimating equation (IEE) framework, which provides a general strategy for quantile identification through signal functions, but their theory is developed for interior quantiles and does not address the structural ill-conditioning that arises when the target quantile moves into the tail.

The common limitation across these approaches is that they treat the causal structure as known. The confounders to be adjusted for, the treatment assignment mechanism, and the absence of unmeasured common causes are taken as given rather than learned. In practice, however, the confounders relevant in the tail may differ from those relevant near the centre, and overlap may hold for central quantiles while failing at the extremes. Knowing which variables require adjustment, and how they are causally related, is precisely the discovery problem. The estimation bottleneck therefore exposes a deeper requirement, which is that one must be able to recover causal structure from extremal data.

3.4 Discovery as learning causal structure in extremes

3.4.1 Standard causal discovery methods

Causal discovery is the task of learning directed acyclic graph (DAG) structure from observational data. It rests on the insight that the conditional independence relationships implied by a causal model constrain the set of compatible graph structures. The classical framework, developed by Spirtes et al. (2000) and Pearl (2009b), establishes that under the causal Markov and faithfulness assumptions, the true causal graph can be identified up to a Markov equivalence class from conditional independence tests on the observed data.

Three families of algorithms have dominated the field. Constraint-based methods, exemplified by the Peter–Clark (PC) algorithm (Spirtes et al., 2000), iteratively test conditional independence constraints to remove edges from a complete graph, then orient remaining edges using orientation rules. Colombo and Maathuis (2014) introduced an order-independent variant that resolves the dependence of PC on variable ordering, improving robustness. The Fast Causal Inference (FCI) algorithm (Spirtes et al., 2000) extends this approach to settings with latent confounders, returning a partial ancestral graph rather than a DAG. Score-based methods, such as greedy equivalence search (GES) (Chickering, 2002), search over the space of Markov equivalence classes by optimising a model selection criterion such as the Bayesian information criterion, providing consistency guarantees under appropriate conditions. Hauser and Bühlmann (2012) extended score-based learning to interventional settings, characterising the Markov equivalence classes that arise when data from multiple experimental conditions are available. Loh and Bühlmann (2014) developed high-dimensional methods for learning linear causal structure via regularised inverse covariance estimation, establishing consistency under sparsity assumptions. Functional causal model approaches exploit asymmetries in the data-generating process to orient edges beyond what conditional independence alone can achieve. For example, Shimizu et al. (2006) showed that non-Gaussianity of the noise terms in a linear structural equation model allows full DAG identification via independent component analysis, and Zheng et al. (2018) formulated the combinatorial graph search as a continuous optimisation problem, enabling gradient-based methods. Peters and Bühlmann

(2014) established identifiability results for Gaussian structural equation models under equal error variance assumptions, providing theoretical foundations for causal structure learning. Peters et al. (2017) provided a comprehensive treatment of causal discovery from a machine learning perspective.

These methods are designed for and calibrated on data from the bulk of the distribution. Their theoretical guarantees rely on regularity conditions such as finite moments, well-behaved densities, and adequate overlap, and these may not hold when the data are heavy-tailed or when the inferential target lies in the extremal regime. The foundational theory for graphical models, developed by Lauritzen (1996), assumes well-behaved probability distributions, so its extension to the extremal setting requires dedicated machinery. In the specific context of Earth system sciences, Runge et al. (2019) proposed a framework for causal inference from time series that combines conditional independence tests with time-delayed relationships, demonstrating that causal networks can be recovered from observational climate data. Their approach operates in the bulk, however, and does not target the extremal regime where tail dependence generates additional complications.

3.4.2 Why methods for the bulk fails in the tail

Heavy-tailed and extremal data violate the assumptions underlying standard discovery methods in specific ways. The fundamental issue is that tail dependence, namely the tendency for variables to simultaneously exceed extreme thresholds, is a distinct phenomenon from the dependence structure in the bulk. Two variables that are conditionally independent at moderate levels can exhibit strong extremal dependence, generating spurious edges in any graph recovered from threshold exceedances. Conversely, variables that appear dependent in the bulk may be extremally independent, causing genuine causal connections to be missed at the extreme levels where they matter most.

Gnecco et al. (2021b) demonstrated this phenomenon explicitly. The classical PC algorithm, applied to data from heavy-tailed elliptical distributions, can be substantially misled because tail dependence induces conditional independence relationships in the extremal limit that do not hold in the bulk, and conversely, bulk independence does not imply extremal independence. Their proposed tail-augmented PC algorithm partially addresses the issue by replacing standard correlation-based tests with extremal dependence measures, but the theoretical analysis is restricted to the class of elliptical distributions.

The characterisation of extremal dependence itself has a substantial literature. Ledford and Tawn (1996) introduced the coefficient of tail dependence η, providing a scalar measure that distinguishes between asymptotic dependence ($\eta = 1$) and asymptotic independence ($\eta < 1$) in the bivariate tail. Ledford and Tawn (1997) extended this framework to model dependence within joint tail regions. Coles et al. (1999) developed a broader toolkit for measuring and interpreting extremal dependence, including the tail dependence coefficient χ and the conditional exceedance measure $\bar{\chi}$. Heffernan and Tawn (2004) proposed a conditional extremes approach that models the conditional distribution of one variable given an extreme value of another, providing a flexible framework that captures both asymptotic dependence and asymptotic independence without requiring a specific parametric copula. The mathematical foundations for this body of work lie in the theory of regular variation, comprehensively treated by Resnick (1987) and later by Resnick (2007). These contributions establish that extremal dependence is a rich and structured phenomenon that cannot be adequately captured by standard correlation or covariance measures, and that dedicated graphical models are therefore needed for the tail.

3.4.3 Extremal graphical models

Engelke and Hitz (2020) introduced the extremal graphical model framework, defining conditional independence in the extremal limit through the tail measure of a multivariate regularly varying distribution. Their construction shows that, analogously to the Gaussian graphical model in the bulk, the extremal conditional independence structure of a random

vector can be represented by an undirected graph, and that the edges of this graph can be estimated from threshold exceedance data. Engelke and Volgushev (2022) developed efficient structure learning algorithms for extremal trees, establishing consistency of the estimated graph under appropriate conditions.

These models are undirected. They capture which variables co-exceed extreme thresholds but do not distinguish cause from effect. For attribution and estimation, this distinction is essential. Knowing that two variables are extremally dependent does not tell us whether one causes the other, or whether a third variable drives both. Directed models for extremes are therefore needed.

3.4.4 Directed models for extremes

Max-linear structural equation models provide a natural directed framework for extremes. In a max-linear model, each variable is the componentwise maximum of its parents' values scaled by positive coefficients, reflecting the max-stability property inherent in extreme value theory. Klüppelberg and Lauritzen (2019) established the connection between max-linear models and Bayesian networks, showing that the recursive structure of a max-linear structural equation model induces a directed acyclic graph that determines the extremal dependence pattern. Gissibl et al. (2018) analysed the tail dependence properties of recursive max-linear models with regularly varying noise terms, characterising how the graph structure manifests in the extremal correlation function. Gissibl et al. (2021) studied identifiability and estimation of recursive max-linear models, showing that the graph can, in principle, be recovered from the tail dependence structure.

Recent work has extended extremal dependence and graphical modelling ideas in several complementary directions, particularly towards causal structure learning in the presence of heavy tails. A first line of research focuses on adapting graphical and constraint-based methods to the extremes. For instance, Tran et al. (2024) developed methods for estimating directed trees from extreme observations using extremal correlation-based tests, while Jiang et al. (2025) proposed a separation-based approach built on transformed-linear structural causal models (drawing on the framework of Cooley and Thibaud 2019), in which partial tail uncorrelatedness enables constraint-based learning of directed acyclic graphs. In a related vein, Engelke and Taeb (2025) extended extremal graphical modelling

to settings with latent variables via convex optimisation. A second strand of work considers structural causal models tailored to heavy-tailed data. Krali (2025) proposed causal discovery methods for heavy-tailed linear structural equation models, whereas Fang et al. (2025) proposed an alternative formulation based on linear programming. These approaches differ in both their modelling assumptions and estimation strategies, but share the goal of recovering causal structure under non-Gaussian, heavy-tailed regimes. Finally, several contributions focus on incorporating confounding and real-world applications. Pasche et al. (2023) developed causal modelling frameworks for heavy-tailed variables in the presence of confounders, while Mhalla et al. (2020) applied tail dependence-based causal analysis to river discharge data in the upper Danube basin, illustrating how these ideas can be deployed in environmental settings.

Taken together, these contributions mark substantial progress, but their scope remains limited in ways that matter for the present thesis. Max-linear approaches impose a specific parametric form on the extremal structural equations. Tree-based methods restrict the graph to a spanning tree, excluding the more general DAGs encountered in applications. Separation-based methods provide an important constraint-based route to directed learning in extremes, but currently they do not address proxy-adjusted orientation under latent common shocks and only recover directions up to Markov equivalence. What remains missing is a directed discovery framework that can recover graph structure from extremal data under weaker prior restrictions and with greater robustness to latent common shocks. The discovery problem therefore remains open precisely where the pipeline most needs it solved. Attribution and estimation require knowing which variables are causes and which are confounders; discovery in the tail must recover that information from the same extreme data on which downstream inference relies.

3.5 Towards an integrated view of causal extremes

The three stages of the pipeline, namely attribution, estimation, and discovery, are connected by more than logical dependency. They share a common mathematical object, which is the extremal dependence structure of the system under study. In attribution, this structure determines whether multiple locations or variables co-exceed their thresholds simul-

taneously, and therefore how joint event probabilities differ between factual and counterfactual worlds. In estimation, it shapes which confounders matter at extreme levels and whether overlap remains plausible for extreme quantiles. In discovery, it is the raw material from which causal graph structure must be recovered through extremal conditional independence or tail asymmetry.

Yet the three literatures have largely developed in isolation. Attribution studies proceed with dependence models chosen for event-probability calculation, often without formal causal identification. Estimation methods assume that the causal structure is already known and concentrate on tail extrapolation within that fixed structure. Discovery methods recover graph topology from extremal data but usually stop short of linking that structure to downstream attribution or treatment-effect estimation. The result is a fragmented literature in which the dependence structure, the causal graph, and the tail model are treated as separate modelling choices even though, in the extreme regime, assumptions made at one level necessarily propagate to the others.

This thesis is positioned against that fragmentation. Chapter 4 studies dependence sensitivity in attribution, showing that extremal dependence assumptions can dominate attribution conclusions even when marginal tail specifications are held fixed. Chapter 5 develops a unified estimation framework that combines causal identification with generalised Pareto tail calibration, avoiding restrictive Fréchet-domain assumptions and sequential estimation strategies. Chapter 6 develops a data-driven framework for causal discovery in multivariate extremes, recovering both skeleton and edge orientation from extremal data without prespecified graph structure. Taken together, these chapters address successive stages of a single inferential problem. The problem is how to draw valid causal conclusions about rare, high-impact events when the data are sparse, the tails are heavy, and the causal structure is unknown.

Chapter 4

Extreme event attribution

This chapter studies how assumptions about tail dependence affect extreme event attribution. It compares three multivariate extreme-value models and shows, through simulations and precipitation applications, that attribution conclusions can change substantially when the dependence structure is misspecified.

A version of this chapter has been submitted for publication under the title *On the Importance of Tail Assumptions in Climate Extreme Event Attribution*. Preprint available at arXiv:2507.14019.

4.1 Introduction

Extreme events in environmental sciences are typically defined as threshold exceedances, i.e., values in the tails of probability distributions whose small probabilities nevertheless make accurate estimation critical for risk assessment. Their increasing frequency and severity in a changing climate has raised major concerns due to potential impacts on societies and ecosystems. The 2022 Global Risks Report identified extreme weather, climate inaction, and biodiversity loss as leading global risks with profound economic implications (World Economic Forum, 2022), while recent studies confirm intensifying climate-related extremes (McPhillips et al., 2018) and directly link them to anthropogenic greenhouse gas emissions (Masson-Delmotte et al., 2021). Against this backdrop, extreme event attribution (EEA) has emerged as a key field for quantifying human influence on extreme weather (van Oldenborgh et al., 2021). By comparing event probabilities under factual (observed anthropogenic forcing) and counterfactual (no anthropogenic forcing) scenarios (Pearl, 2009a; Shadish et al., 2002; Hannart et al., 2016), EEA enables a formal assessment of whether human actions have made particular events more likely or more severe.

The attribution results derived from this causal framework are increasingly used beyond the statistical literature, including in public communication about climate change and in assessments of climate-related losses and damages (Noy et al., 2023). In such settings, attribution statements are often summarized as probabilities or risk ratios, while the underlying modeling assumptions required to estimate these quantities remain implicit. Previous studies have highlighted that attribution conclusions can be sensitive to climate model choice and experimental design (Hauser et al., 2017). Moreover, attribution results depend critically on how extreme events are defined, including the choice of climate variable or index used to characterize the event (Lanet et al., 2024), as well as the spatial and temporal scales that delineate the event (Kirchmeier-Young et al., 2019). This raises the question of how sensitive reported attribution probabilities are to methodological choices more broadly. In this chapter, we focus specifically on those related to the statistical modeling of rare and extreme events.

A wide range of statistical approaches has been developed to estimate these probabilities. Non-parametric methods model exceedance counts as binomial processes, deriving risk ratios from factual and counterfactual comparisons (Paciorek et al., 2018), with Paciorek et al. (2018) also providing a framework for uncertainty quantification. Yet such methods become computationally demanding for rare, high-threshold events (Naveau et al., 2020b). Parametric approaches, though assumption-driven, generally yield more efficient inference for small probabilities. Extreme-value theory (EVT) is particularly well suited to EEA as it characterizes tail behavior and allows extrapolation to unprecedented extremes. Applications include rapid attribution of precipitation maxima in Louisiana using the generalized extreme-value distribution (Van Der Wiel et al., 2017), frameworks for non-stationary records to assess trends in record-breaking precipitation (Gonzalez et al., 2025), and attribution of daily U.S. precipitation extremes via the extended generalized Pareto distribution (Nanditha et al., 2024). However, these univariate or location-specific methods often neglect dependence across sites, which may differ across factual and counterfactual climate (Kirchmeier-Young et al., 2019). To overcome this limitation, Kiriliouk and Naveau (2020) combined multivariate EVT, specifically, the multivariate generalized Pareto distribution (mGPD), with Pearl's counterfactual causation theory to extend attribution to high-dimensional extremes.

Following this line, we study high-threshold multivariate exceedance events, where at least one component of a multivariate vector (e.g., precipitation at different locations) exceeds a specified threshold. Our goal is to assess how assumptions about marginal and joint tail behavior affect causal metrics by comparing three statistical models. The mGPD provides an asymptotically justified approach but assumes persistent asymptotic dependence, which may not hold in environmental contexts. The exponential factor copula model (eFCM) of Castro-Camilo and Huser (2020) relaxes this by allowing extremal dependence to decay with event severity through a latent factor, while the Huser–Wadsworth (HW) model (Huser and Wadsworth, 2019) offers even greater flexibility to capture a wide range of dependence structures. Both models, originally designed for spatial extremes, are here adapted to multivariate data.

To evaluate these approaches, we conduct two simulation studies. The first examines model misspecification by altering margins or dependence; the second benchmarks the mGPD and eFCM against data generated from the HW model. In addition, we consider two real-data applications: European precipitation from the CNRM climate model (Kiriliouk and Naveau, 2020) and U.S. precipitation from ACCESS-CM2 (Nanditha et al., 2024). These studies are designed to assess how different tail modeling choices influence attribution metrics in practice.

The remainder of this chapter is structured as follows. Section 4.2 introduces the methodological framework and the three modeling approaches. Section 4.3 presents the simulation studies, Section 4.4 reports the data applications, and Section 4.5 summarizes the main findings and implications. The C++ and R code for implementing the statistical models and reproducing all results is available at `https://github.com/MengranLi-git/tail-eea`. The eFCM model can also be accessed directly through the R package `eFCM` (Li and Castro-Camilo, 2025).

4.2 Methodological framework

In this section we define extreme events within a counterfactual attribution framework, introduce the causal estimands used in EEA, describe three candidate multivariate tail models and their dependence properties, and explain how we estimate the key exceedance probabilities that drive attribution metrics. A consolidated list of recurring symbols used in this and subsequent chapters is provided in the front matter.

4.2.1 Attribution setup and causal estimands

This section adapts the counterfactual attribution notation and PN formula introduced in Section 2.1.1.2, especially Eqs. (2.7)–(2.8), to the multivariate EEA setting. Throughout the thesis, p_1/p_0 is termed the risk ratio (RR), while $(p_0 - p_1)/p_1$ is termed the attribution ratio (AR). Let Y represent the binary indicator of an extreme event, assumed to occur when a climatological index exceeds a high threshold v. Let $T \in \{0, 1\}$ be the treatment effect, encoding the absence ($T = 0$, counterfactual) or presence ($T = 1$, factual) of anthropogenic forcing. We denote by $Y(1)$ and $Y(0)$ the corresponding potential outcomes. In multivariate settings, let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ collect climate variables (e.g., precipitation over d locations) and let $\boldsymbol{w} = (w_1, \ldots, w_d)^\top$ denote nonnegative weights. We consider high-threshold events of the form $Y = \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{X} > v\}$. Let $\boldsymbol{X}^{(t)}$ denote the multivariate climate vector under world $t \in \{0, 1\}$. Then, the interest is on

$$p_t = \Pr\{\boldsymbol{w}^\top \boldsymbol{X}^{(t)} > v\}, \qquad t \in \{0, 1\}. \tag{4.1}$$

Following the counterfactual causal framework of Pearl (2009a), estimation of causal effects rests on four standard assumptions: (i) *consistency* (the observed Y equals the potential outcome under the received treatment); (ii) *no interference*/stable unit treatment value; (iii) *unconfoundedness*, $(Y(0), Y(1)) \perp\!\!\!\perp T \mid \boldsymbol{X}$ for observed covariates $\boldsymbol{X}$; and (iv) *positivity*, $0 < \Pr(T = 1 \mid \boldsymbol{X}) < 1$ for all $\boldsymbol{X}$. Within this framework, causal notions such as the probabilities of necessary, sufficient, and necessary and sufficient causation, can be expressed in terms of p_0 and p_1 as

$$\mathrm{PN} = \max\left(1 - \frac{p_0}{p_1}, 0\right), \qquad \mathrm{PS} = \max\left(1 - \frac{1 - p_0}{1 - p_1}, 0\right), \qquad \mathrm{PNS} = \max(p_1 - p_0, 0).$$

Additionally, two attribution metrics commonly used in EEA are the risk and attribution ratios, defined as

$$\mathrm{RR} = \frac{p_1}{p_0}, \qquad \mathrm{AR} = \frac{p_0 - p_1}{p_1}.$$

When $p_0 < p_1$, AR is negative, indicating risk enhancement under anthropogenic influence; when $p_1 < p_0$, AR is positive, indicating risk reduction and PN = 0. Our analyses focus on AR (with PN reported in Appendix A.5).

4.2.2 Multivariate generalized Pareto distribution (mGPD)

The mGPD is the asymptotic model for multivariate threshold exceedances. Following Rootzén and Tajvidi (2006); Rootzén et al. (2018), define the generalized Pareto random vector via the construction

$$\boldsymbol{Z}^* \overset{d}{=} E + \boldsymbol{T} - \max_{1\le j\le d} T_j, \quad \boldsymbol{Z} \overset{d}{=} \frac{\boldsymbol{\sigma}}{\boldsymbol{\gamma}}\{\exp(\boldsymbol{\gamma}\boldsymbol{Z}^*) - 1\},$$

with $E \sim \mathrm{Exp}(1)$ independent of $\boldsymbol{T} = (T_1, \ldots, T_d)^\top$, $\boldsymbol{\sigma} > \boldsymbol{0}$, and $\boldsymbol{\gamma} \in \mathbb{R}^d$. If $\boldsymbol{T}$ has density $f_{\boldsymbol{T}}$, then

$$f_{\boldsymbol{Z}^*}(\boldsymbol{z}) = \frac{\mathbb{1}\{\max(\boldsymbol{z}) > 0\}}{\exp\{\max(\boldsymbol{z})\}} \int_0^\infty f_{\boldsymbol{T}}(\boldsymbol{z} + \log t)\, t^{-1}\, \mathrm{d}t, \tag{4.2}$$

and

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{\boldsymbol{Z}^*}(\mathbb{1}/\boldsymbol{\gamma}\log(1 + \boldsymbol{\gamma}\boldsymbol{z}/\boldsymbol{\sigma})) \prod_{j=1}^d \frac{1}{\sigma_j + \gamma_j z_j}.$$

We use generators with independent reverse-exponential components,

$$f_{\boldsymbol{T}}(\boldsymbol{z}) = \prod_{j=1}^d \alpha_j \exp\{\alpha_j(z_j + \beta_j)\},$$

where $z_j < -\beta_j$, $\alpha_j > 0$, and $\beta_j \in \mathbb{R}$. Under this assumption, the density in (4.2) becomes

$$f_{\boldsymbol{Z}^*}(\boldsymbol{z}) = \frac{\exp\left\{-\max(\boldsymbol{z}) - \max(\boldsymbol{z} + \boldsymbol{\beta}) \sum_{j=1}^d \alpha_j\right\}}{\sum_{j=1}^d \alpha_j} \prod_{j=1}^d \alpha_j (\exp\left\{\alpha_j(z_j + \beta_j)\right\}).$$

Although the marginal distributions of the mGPD are not necessarily univariate GPDs, this result holds when considering conditional marginal distributions. Specifically, the j-th conditional marginal survival function corresponds to the survival function of the univariate GPD H, such that

$$\Pr[Z_j > z \mid Z_j > 0] = \bar{H}(z; \sigma_j, \gamma_j), \tag{4.3}$$

for any $z > 0$ and $j \in \{1, ..., d\}$, where

$$\bar{H}(z; \sigma_j, \gamma_j) = \left(1 + \gamma_j \frac{z}{\sigma_j}\right)_+^{-1/\gamma_j}, \quad \sigma_j > 0, \ \gamma_j \in \mathbb{R}. \tag{4.4}$$

Assuming that $\gamma_j \equiv \gamma$, i.e., the shape parameter is the same for all locations, (4.3) can be computed using linear projections, as described in Rootzén et al. (2018). Specifically, if $\boldsymbol{Z} \sim \text{mGPD}(\boldsymbol{T}, \boldsymbol{\sigma}, \gamma\mathbf{1})$, then for any nonnegative weights $\boldsymbol{w} = (w_1, \cdots, w_d)^\top$ such that $\Pr(\boldsymbol{w}^\top \boldsymbol{Z} > 0) > 0$, the conditional distribution of $\boldsymbol{w}^\top \boldsymbol{Z}$ given $\boldsymbol{w}^\top \boldsymbol{Z} > 0$ is

$$[\boldsymbol{w}^\top \boldsymbol{Z} \mid \boldsymbol{w}^\top \boldsymbol{Z} > 0] \sim \text{GPD}(\boldsymbol{w}^\top \boldsymbol{\sigma}, \gamma). \tag{4.5}$$

Assuming tail homogeneity across the study region requires prior analysis to assess its suitability in practice. While the result in (4.5) offers computational efficiency for marginal probability estimation, the mGPD model remains computationally demanding overall. Additionally, its underlying assumption of persistent strong tail dependence at increasingly extreme levels may not hold in many environmental applications, where evidence often points to weakening dependence in the tails. These considerations motivate the use of the two alternative subasymptotic models introduced below.

4.2.3 Exponential factor copula model (eFCM)

Castro-Camilo and Huser (2020) propose a censored local likelihood approach for a subasymptotic spatial process W, stochastically defined as the sum of a Gaussian process and a common exponential factor. Specifically

$$W(s) = Z(s) + V, \quad s \in \boldsymbol{S} \subset \mathbb{R}^2, \tag{4.6}$$

where $\boldsymbol{S}$ denotes the study area, $Z(s)$ is a zero-mean Gaussian process with a stationary correlation function $\rho(h)$ (with $h = \|s_1 - s_2\|, s_1, s_2 \in \boldsymbol{S}$), and $V \geq 0$ is an exponentially distributed random variable with rate parameter $\lambda > 0$, independent of both $Z(s)$ and the spatial location $s \in \boldsymbol{S}$. Thus, the process W, which falls within the class of asymptotically dependent models, can be viewed as a Gaussian location mixture, i.e., a Gaussian process with a random (exponentially distributed) mean. Letting $Z_j = Z(s_j), j = 1, ..., d$, the random vector $\boldsymbol{Z} = (Z_1, ..., Z_d)^\top$ follows a multivariate normal distribution, $\boldsymbol{Z} \sim \Phi(\cdot; \Sigma_Z)$, where the covariance matrix Σ_Z is determined by the correlation function $\rho(h)$. Here, we specify $\rho(h)$ as the exponential correlation function $\rho(h) = \exp\{-h/\delta\}$, with δ denoting the range parameter. We similarly define $\boldsymbol{W} = (W_1, ..., W_d)^\top$, where $W_j = Z_j + V$, for $j = 1, ..., d$, and $V \sim \exp(\lambda)$ is independent of $\boldsymbol{Z}$.

The inclusion of the common factor V restricts the model in (4.6) to homogeneous regions, akin to the tail homogeneity assumption underlying the result in (4.5). As a result, preliminary checks for homogeneity across locations are necessary before applying the model. The censored local likelihood approach proposed by Castro-Camilo and Huser (2020) addresses this limitation by partitioning the study area into smaller, approximately homogeneous regions, allowing the model to be applied more flexibly across larger and more heterogeneous spatial domains.

4.2.4 Huser–Wadsworth (HW) model

Huser and Wadsworth (2019) propose a flexible stationary model that captures both asymptotic dependence and asymptotic independence in spatial extremes, allowing a smooth transition between these two classes without requiring pre-specification. The HW model is defined as

$$X(\boldsymbol{s}) = R^\delta W(\boldsymbol{s})^{1-\delta}, \quad \delta \in [0, 1], \ \boldsymbol{s} \in \boldsymbol{S} \subset \mathbb{R}^2, \tag{4.7}$$

where $W(\boldsymbol{s})$ is a stationary spatial process with standard Pareto margins, R is an independent standard Pareto random variable, and δ controls the strength of extremal dependence. When $\delta > 1/2$, R^δ dominates, leading to asymptotic dependence; when $\delta < 1/2$, $W(\boldsymbol{s})^{1-\delta}$ dominates, resulting in asymptotic independence. Inference is performed using a censored

likelihood approach based on a copula construction. The choice of model for $W(\boldsymbol{s})$ can vary depending on the application, with the authors exploring options such as Gaussian processes, inverted extreme value logistic model and inverted max-stable processes. Here, we will use Gaussian processes to model the dependence structure. As with the model in (4.6), the HW model can be adapted to the multivariate setting by replacing the process $W(s)$ with an asymptotically independent random vector $(W_1, \ldots, W_d)^\top$.

4.2.5 Extremal-dependence summaries

Section 2.2.3 introduces the generic tail-dependence coefficients in Eqs. (2.24)–(2.25). In the present chapter, these coefficients provide a common basis for comparing the dependence structures implied by the three tail models. To understand how different climate models represent joint extremes, it is important to examine the extremal dependence structures they imply. This can provide insight into potential differences in estimated causal probabilities, as introduced in Section 4.2.1. In this section, we characterize extremal dependence using two measures: the bivariate tail dependence coefficient $\chi \in [0, 1]$ and the conditional exceedance probability $\chi(u)$. The coefficient χ captures the strength of dependence in the asymptotic (limiting) tail, while $\chi(u)$ provides a finite-level analogue, measuring the probability of joint exceedances at high but non-asymptotic thresholds. For vector components Y_1 and Y_2, these coefficients are defined as

$$\chi = \lim_{u \to 1} \chi(u) := \lim_{u \to 1} \Pr(F_{Y_1}(Y_1) > u \mid F_{Y_2}(Y_2) > u).$$

For the multivariate generalized Pareto distribution (mGPD) with independent reverse exponential generators, the bivariate tail dependence coefficient χ_{mGPD} admits a closed-form expression (see the Supplementary Material of Kiriliouk et al. 2019):

$$\chi_{\text{mGPD}} = 1 - \left(\frac{1 - \alpha_{(1)}^{-1}}{1 + \alpha_{(2)}^{-1}} \right)^{1+\alpha_{(2)}} \frac{\alpha_{(1)}}{\alpha_{(2)}} \frac{1}{1 + \alpha_1 + \alpha_2},$$

where α_1 and α_2 are the positive rate parameters of the two independent reverse-exponential generator components introduced above, and $\alpha_{(1)} = \max(\alpha_1, \alpha_2)$ and $\alpha_{(2)} = \min(\alpha_1, \alpha_2)$. The corresponding $\chi_{\text{mGPD}}(u)$ does not have a closed-form expression for finite u but can be estimated numerically, e.g., via Monte Carlo or quadrature.

For the eFCM, the tail dependence coefficient is given by

$$\chi_{\text{eFCM}} = 2\left\{1 - \Phi\left(\lambda\sqrt{\frac{(1-\rho_{12})}{2}}\right)\right\},$$

where λ is the rate parameter and ρ_{12} is the correlation (in our case, derived from an exponential correlation model). The corresponding $\chi_{\text{eFCM}}(u)$ can be computed as

$$\chi_{\text{eFCM}}(u) = \frac{1 - 2u + \Phi_2(z(u), z(u); \rho) - 2\exp\{\lambda^2/2 - \lambda z(u)\Phi_2(q; 0, \boldsymbol{\Omega})\}}{1-u},$$

where $z(u) = F_1^{-1}(u;\lambda)$ is the marginal quantile function, $\Phi_2(\cdot,\cdot,\rho)$ is the bivariate standard normal CDF with correlation ρ, $q = \lambda(1-\rho)$, and $\boldsymbol{\Omega}$ is the correlation matrix.

For the HW model, closed-form expressions for χ_{HW} and $\chi_{\text{HW}}(u)$ are generally not available, but both can be numerically approximated. If $\delta > 1/2$, the process is asymptotically dependent, and the limiting coefficient is

$$\chi_{\text{HW}} = \mathbb{E}\left\{\min(W_1, W_2)^{(1-\delta)/\delta}\right\}\frac{2\delta - 1}{\delta}$$

Moreover, if $\delta \neq 1/2$ and u is near 1, we have

$$\chi_{\text{HW}}(u) = \frac{1 - 2u + \displaystyle\int_0^{\log w(u)/\delta} F_{\tilde{W}_1,\tilde{W}_2}\left(\frac{\log w(u) - \delta r}{1-\delta}, \frac{\log w(u) - \delta r}{1-\delta}\right) e^{-r}\,\mathrm{d}r}{1-u},$$

where $w(u) = F_1^{-1}(u)$, is the marginal quantile function of the HW model, and $F_{\tilde{W}_1,\tilde{W}_2}$ is the joint distribution of the log-transformed components $\tilde{W}_1 = \log W_1$ and $\tilde{W}_2 = \log W_2$. At the boundary case $\delta = 1/2$, the process is asymptotically independent, so $\chi_{\text{HW}} = 0$. Nevertheless, the decay of $\chi_{\text{HW}}(u)$ is unusually slow:

$$\chi_{\text{HW}}(u) \sim \frac{1-u}{\log\{1/(1-u)\}}, \quad \text{as } u \to 1,$$

i.e., $\chi_{\text{HW}}(u)$ decays to zero more slowly than any power of $1-u$. Step-by-step derivations of $\chi_{\text{eFCM}}(u)$ and $\chi_{\text{HW}}(u)$ are provided in Appendix A.2.

4.2.6 Estimating p_0 and p_1 and choosing the projection

In this section, we describe how to compute the extreme probabilities in (4.1) using the three statistical models introduced in Sections 4.2.2–4.2.4. We also explain the construction of the nonnegative weight vector $\boldsymbol{w} = (w_1, \cdots, w_d)^\top$ used in these calculations.

Let $\boldsymbol{u} = (u_1, \cdots, u_d)^\top$ be a vector of high marginal thresholds, and let $\boldsymbol{w} = (w_1, \cdots, w_d)^\top$ be the vector of nonnegative weights such that $v > \boldsymbol{w}^\top \boldsymbol{u}$. For $t \in \{0, 1\}$, $\boldsymbol{u}^{(t)}$ denotes the corresponding vector of componentwise intermediate thresholds in world t.

Under the mGPD model, and using the projection property in (4.5), the probability of a weighted exceedance can be approximated by

$$\begin{aligned} \Pr[\boldsymbol{w}^\top \boldsymbol{X} > v] &= \Pr[\boldsymbol{w}^\top \boldsymbol{X} > \boldsymbol{w}^\top \boldsymbol{u}] \cdot \Pr[\boldsymbol{w}^\top (\boldsymbol{X} - \boldsymbol{u}) > v - \boldsymbol{w}^\top \boldsymbol{u} \mid \boldsymbol{w}^\top (\boldsymbol{X} - \boldsymbol{u}) > 0] \\ &\approx \Pr[\boldsymbol{w}^\top \boldsymbol{X} > \boldsymbol{w}^\top \boldsymbol{u}] \cdot \bar{H}(v - \boldsymbol{w}^\top \boldsymbol{u}; \boldsymbol{w}^\top \boldsymbol{\sigma}, \gamma), \end{aligned} \tag{4.8}$$

where $\bar{H}(\cdot; \boldsymbol{w}^\top \boldsymbol{\sigma}, \gamma)$ is the survival function of the univariate GPD with scale parameter $\boldsymbol{w}^\top \boldsymbol{\sigma}$ and common shape parameter γ (see (4.4)). The exceedance probability in the first term of (4.8) can be estimated empirically as

$$\widehat{p}^{\,\mathrm{emp}}(\boldsymbol{w}^\top \boldsymbol{u}^{(t)}; \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{X}_i^{(t)} > \boldsymbol{w}^\top \boldsymbol{u}^{(t)}\}, \quad t \in \{0, 1\}.$$

If $v \le \boldsymbol{w}^\top \boldsymbol{u}$, then the event lies within the bulk of the distribution, and we estimate the probability $\Pr[\boldsymbol{w}^\top \boldsymbol{X} > v]$ directly

$$\widehat{p}^{\,\mathrm{emp}}(v; \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{X}_i^{(t)} > v\}, \quad t \in \{0, 1\}$$

Putting both cases together, we define the mGPD estimator for $p_t = \Pr[\boldsymbol{w}^\top \boldsymbol{X}^{(t)} > v]$, $t \in \{0, 1\}$ as

$$\widehat{p}_t(v; \boldsymbol{w}) = \begin{cases} \widehat{p}_t^{\,\mathrm{emp}}(v; \boldsymbol{w}), & \text{if } v \le \boldsymbol{w}^\top \boldsymbol{u}^{(t)}, \\ \widehat{p}_t^{\,\mathrm{emp}}(\boldsymbol{w}^\top \boldsymbol{u}^{(t)}; \boldsymbol{w}) \bar{H}[v - \boldsymbol{w}^\top \boldsymbol{u}^{(t)}; \boldsymbol{w}^\top \hat{\boldsymbol{\sigma}}, \hat{\gamma}], & \text{if } v > \boldsymbol{w}^\top \boldsymbol{u}^{(t)}. \end{cases}$$

For the eFCM and HW models, the probabilities p_t are estimated by Monte Carlo integration over samples $\boldsymbol{X}_i'^{(t)}$ drawn from the fitted models:

$$\widehat{p}_t(v; \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{X}_i'^{(t)} > v\}, \quad t \in \{0, 1\}.$$

To construct the nonnegative weight vector $\boldsymbol{w} = (w_1, \ldots, w_d)^\top$, we follow the approach of Kiriliouk and Naveau (2020), which proposes maximizing measures of causal attribution, specifically, the probability of necessary causation or the attribution ratio. The key idea is to optimize over linear projections $\boldsymbol{w}^\top \boldsymbol{X}$ to identify the direction in the multivariate space that is most sensitive to the influence of external forcing. This projection can be interpreted as a near-sufficient statistic for distinguishing factual and counterfactual extremes. It enhances statistical efficiency by concentrating on the features that most strongly capture changes in tail probabilities due to forcing. Instead of fixing $\boldsymbol{w}$ a priori, the data-driven choice of $\boldsymbol{w}$ allows us to adaptively focus on the dimensions of the system most affected by anthropogenic influence. To this end, we treat the extreme probabilities as a function of $\boldsymbol{w}$, i.e.,

$$p_t(\boldsymbol{w}) := \Pr(\boldsymbol{w}^\top X^{(t)} > v).$$

Our goal is to find the weight vector $\boldsymbol{w}^*$ that maximizes either the PN or AR. These are defined as:

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \mathrm{PN}(\boldsymbol{w}) := \arg\max_{\boldsymbol{w}} \left(1 - \frac{p_0(\boldsymbol{w})}{p_1(\boldsymbol{w})}\right)_+$$

or

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \mathrm{AR}(\boldsymbol{w}) := \arg\max_{\boldsymbol{w}} \frac{p_0(\boldsymbol{w}) - p_1(\boldsymbol{w})}{p_1(\boldsymbol{w})}$$

The motivation for defining $\boldsymbol{w}^*$ as the maximizer of PN or AR stems from the goal of identifying the most informative linear combination of variables that reveals the effect of anthropogenic forcing on the probability of extreme events. Specifically, this approach: (i) identifies the linear projection most influenced by the forcing; (ii) maximizes contrast between factual and counterfactual exceedance probabilities; (iii) aligns with the counterfactual framework for event attribution; and (iv) supports interpretable attribution statements about which features are driving the observed changes.

4.3 Simulation study

To investigate the impact of tail assumptions on extreme event attribution, we conduct two simulation studies comparing the three multivariate extreme value models introduced in Section 4.2. The first scenario compares the effects of one marginal distortion and one dependence misspecification under an HW data-generating model. The second scenario uses the HW model to simulate data under different asymptotic regimes, evaluating the performance of the mGPD and eFCM models. Within the specified HW designs and parameter values, the eFCM yields lower estimation error than the mGPD.

4.3.1 Simulation scenario 1

Accurate specification of both the marginal distributions and the dependence structure is essential in multivariate modeling. However, in practice, it is often difficult to get both right—a challenge we also encountered in our data applications. The comparison includes a correctly specified baseline with right dependence and right margins (RD–RM). The remaining cases use right dependence and wrong margins (RD–WM), wrong dependence and right margins (WD–RM), and wrong dependence and wrong margins (WD–WM).

We chose the HW model as the true data-generating process because its stochastic representation naturally separates the marginal distributions from the dependence structure. We generate $n = 1,000$ replicates of the HW model at $d = 2$ locations, assuming a Gaussian process for W with an exponential covariance function with range parameter 0.8. We consider two values of δ, namely $\delta = 0.3$ (asymptotic independence) and $\delta = 0.7$ (asymptotic dependence). Let $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ denote the bivariate simulated vectors, with $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in})^\top$, $i = 1, 2$. Our target is to estimate $p = \Pr(w_1 X_1 + w_2 X_2 > v_p)$, where v_p is the p-quantile of the weighted sum $w_1 X_1 + w_2 X_2$, for $p = 0.01, 0.02, 0.05$ and $w_1 = w_2 = 1/2$. We repeat the entire simulation procedure MC $= 1,000$ times to assess model performance.

4.3.1.1 Sub-scenario 1.1: misspecified margins, correct dependence

In this setting, we assume the correct dependence structure ("right dependence" or RD) for W but intentionally misspecify the margins ("wrong margins" or WM). Specifically, we introduce marginal distortion by applying the transformation $\boldsymbol{X}_i^e = F_p^{-1}\{F_e(\boldsymbol{X}_i)\}$, where F_e is the unit exponential distribution and F_p^{-1} is the inverse of the standard Pareto distribution. We then fit the HW model to the transformed data $(\boldsymbol{X}_1^e, \boldsymbol{X}_2^e)$, using the correct dependence structure (i.e., the original Gaussian process with true covariance). From the fitted model, we generate $m = 100,000$ samples $(\tilde{X}_{1,1}, \ldots, \tilde{X}_{1,m}, \tilde{X}_{2,1}, \ldots, \tilde{X}_{2,m})^\top$ and estimate p using the empirical estimator $\hat{p} = \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}\{w_1\tilde{X}_{1,k} + w_2\tilde{X}_{2,k} > v_p\}$ for each value of p.

4.3.1.2 Sub-scenario 1.2: misspecified dependence, correct margins

Here, the margins are correctly specified as standard Pareto, but we misspecify the dependence structure by using the inverted extreme value logistic (IEVL) model for W, which results in incorrect spatial dependence modeling; see Figure A.2 in Appendix A.3 for a comparison of the Gaussian and IEVL copulae. As in Sub-scenario 1.1, we use the fitted model to simulate $m = 100,000$ replicates and estimate p via its empirical estimator.

Figure 4.1 shows the bias of $\hat{p}$ for all four scenarios across the different values of p, using boxplots based on MC $= 1,000$ Monte Carlo replicates. The four scenarios are the correctly specified baseline RD–RM, the marginal-only misspecification RD–WM (Sub-scenario 1.1), the dependence-only misspecification WD–RM (Sub-scenario 1.2), and the both-misspecified scenario WD–WM.

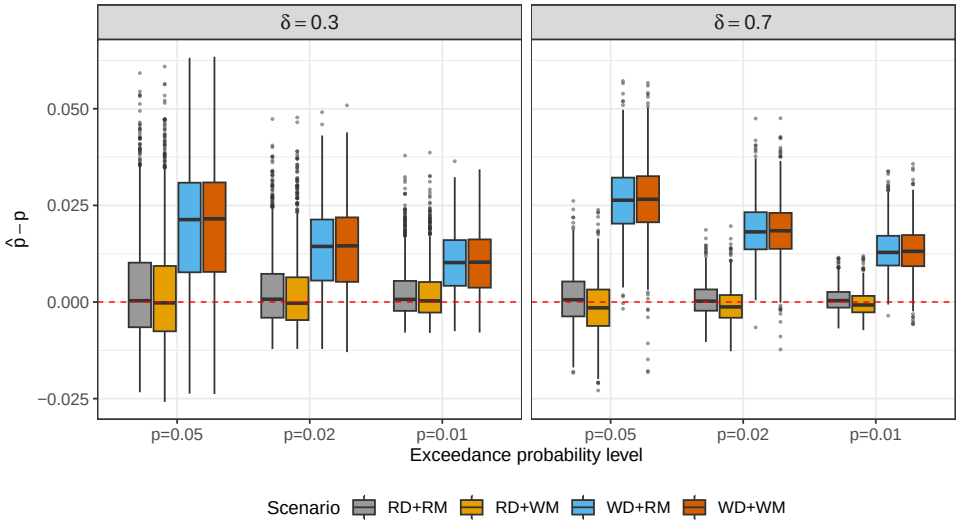

Figure 4.1: Bias $\hat{p} - p$ for target probabilities $p \in \{0.01, 0.02, 0.05\}$, with $\delta = 0.3$ (left) and $\delta = 0.7$ (right), under the full 2×2 design of Section 4.3.1: correct dependence and correct margins (RD–RM, baseline); correct dependence and wrong margins (RD–WM, Sub-scenario 1.1); wrong dependence and correct margins (WD–RM, Sub-scenario 1.2); wrong dependence and wrong margins (WD–WM). For $m = 100{,}000$ draws from a fitted model, $\hat{p} = m^{-1} \sum_{k=1}^{m} \mathbb{1}\{(\tilde{X}_{1k} + \tilde{X}_{2k})/2 > v_p\}$ estimates the data-generating tail probability p. Each boxplot contains 1000 Monte Carlo replicates; the red dashed line marks zero bias.

The baseline RD–RM and marginal-only RD–WM scenarios are both centred close to zero bias across the values of p and δ. The similarity of these two cases reflects that the monotone marginal distortion in Sub-scenario 1.1 preserves the ranks used to fit the Gaussian copula. By contrast, the dependence-only WD–RM scenario produces clear positive bias, which is larger for higher p and δ. The both-misspecified WD–WM scenario closely follows WD–RM, further indicating that dependence misspecification has a comparatively large effect under this simulation design.

4.3.2 Simulation scenario 2

Having compared marginal and dependence misspecification in the preceding HW-based design, we now compare the performance of the mGPD and eFCM models in estimating the tail probabilities $p = \Pr(w_1 X_1 + w_2 X_2 > v_p)$. As in the first simulation, we use the HW model as the data-generating process due to its flexibility in capturing various asymptotic dependence regimes.

We generate $n = 1,000$ replicates from the HW model, again assuming a Gaussian process for W with an exponential covariance function and range parameter 0.8. To explore a broader spectrum of asymptotic dependence, we consider five values of the dependence parameter: $\delta = 0.3,\ 0.4,\ 0.6,\ 0.7,\ 0.8$. For each setting, we repeat the simulation procedure MC = 1,000 times to assess model performance.

Figure 4.2 shows estimates of the conditional probability $\chi(u)$ for $u \in [0.5, 1]$ obtained from both models, with the true $\chi(u)$ curve from the HW model included for comparison. The results for $\delta = 0.8$ were omitted as they are almost identical to those of $\delta = 0.7$. As expected, the mGPD displays its characteristic constant $\chi(u)$ profile, which leads to overestimation of extremal dependence (u close to 1) when $\delta < 0.5$ (cases of asymptotic independence) and fails to reflect the decay in dependence typical of subasymptotic regimes. In contrast, the eFCM provides a much closer match to the true $\chi(u)$ across all δ values, successfully capturing the changing dependence structure.

Figure 4.3 presents estimates of the tail probabilities $p = 0.01,\ 0.02,\ 0.05$ for each value of δ. Within these HW settings, the eFCM estimates are closer to the generated tail probabilities and have lower error, especially for the smaller values of δ considered. The mGPD estimates show greater bias and variability in those same settings. Table 4.1 reports the root mean square error (RMSE) for estimates of $\hat{p}$. Across the HW settings and parameter values considered here, the eFCM has lower RMSE than the mGPD.

Table 4.1: Root mean squared error of the mGPD and eFCM estimators of the data-generating tail probability p in the second simulation scenario (Section 4.3.2), for $p \in \{0.01, 0.02, 0.05\}$ and $\delta \in \{0.3, 0.4, 0.6, 0.7, 0.8\}$. With replicate estimates $\hat{p}_b$ and $B = 1000$, RMSE $= \{B^{-1} \sum_{b=1}^{B} (\hat{p}_b - p)^2\}^{1/2}$. Bold identifies the smaller RMSE for each (p, δ) combination.

		δ				
Model	p	0.3	0.4	0.6	0.7	0.8
eFCM	0.05	**0.0017**	**0.0012**	**0.0008**	**0.0007**	**0.0007**
mGPD		0.0049	0.0043	0.0036	0.0035	0.0038
eFCM	0.02	**0.0008**	**0.0006**	**0.0005**	**0.0004**	**0.0004**
mGPD		0.0030	0.0031	0.0027	0.0026	0.0025
eFCM	0.01	**0.0005**	**0.0004**	**0.0003**	**0.0003**	**0.0003**
mGPD		0.0025	0.0024	0.0021	0.0022	0.0022

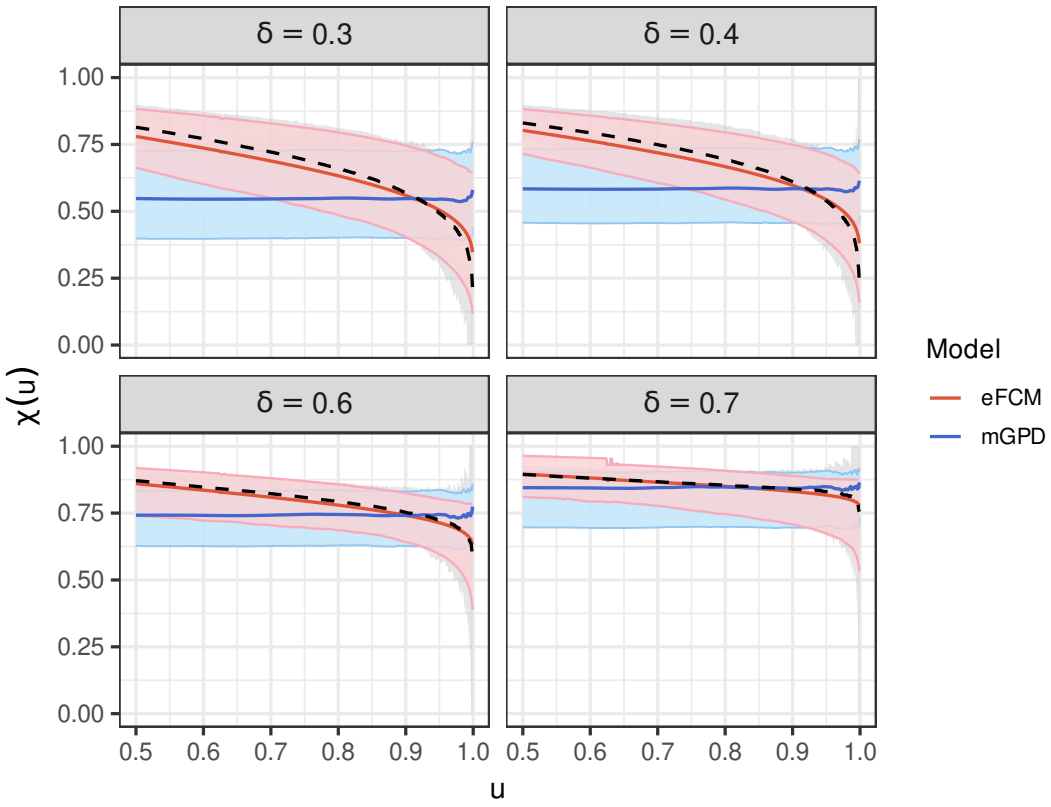

Figure 4.2: Estimated and data-generating values of the conditional exceedance probability $\chi(u) = \Pr\{F_1(X_1) > u \mid F_2(X_2) > u\}$ for the second simulation study (Section 4.3.2), over $u \in [0.5, 1]$ and $\delta \in \{0.3, 0.4, 0.6, 0.7\}$. The black dashed line is the HW data-generating curve, while the red and blue lines are the eFCM and mGPD estimates. The grey, red, and blue shaded regions are the corresponding 95% pointwise Monte Carlo intervals across 1000 replicates.

4.4 Data applications

To assess the extent to which human activities have contributed to extreme precipitation events, we combine the causal inference framework for event attribution (Section 4.2.1) with the extreme value modeling techniques introduced in Sections 4.2.2 to 4.2.4. A central challenge in this setting is the inherent unobservability of the counterfactual world. To circumvent this, we use climate model simulations to generate data representing both the factual and counterfactual scenarios.

We focus on two applications concerning precipitation extremes. The first dataset contains simulations from the CNRM global climate model developed by Météo-France, part of the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al. 2016; Voldoire et al. 2019), providing precipitation outputs over Europe under both factual and counterfactual conditions. The second dataset offers similar simulations across the con-

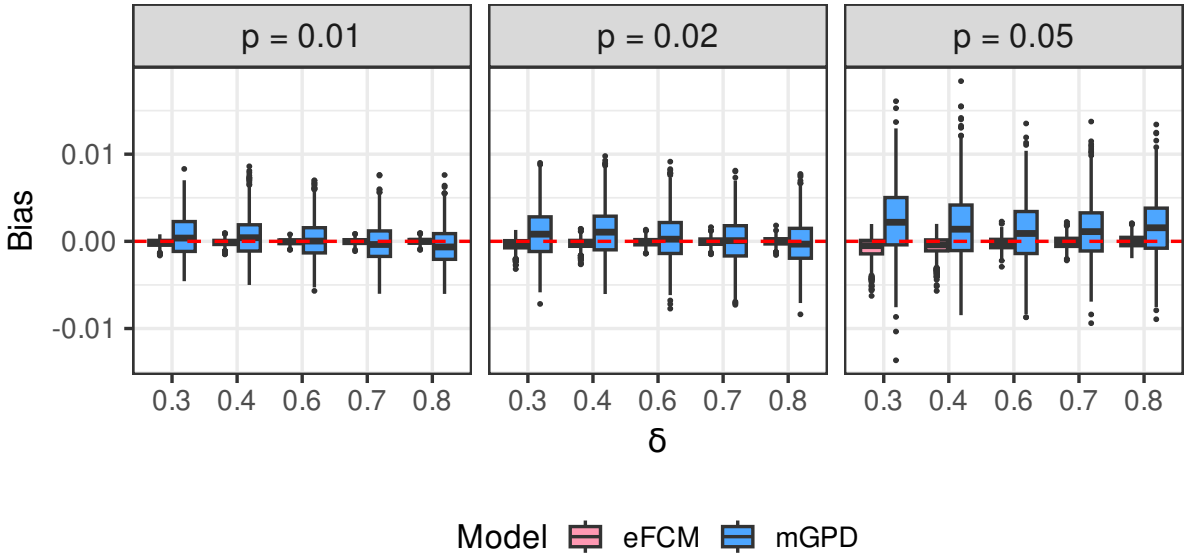

Figure 4.3: Bias $\hat{p} - p$ in the second simulation scenario (Section 4.3.2) for the mGPD and eFCM estimators, with $p \in \{0.01, 0.02, 0.05\}$ and $\delta \in \{0.3, 0.4, 0.6, 0.7, 0.8\}$. Here $\hat{p} = m^{-1} \sum_{k=1}^{m} \mathbb{1}\{(\tilde{X}_{1k} + \tilde{X}_{2k})/2 > v_p\}$ with $m = 100{,}000$ fitted-model draws, and each boxplot summarizes 1000 Monte Carlo replicates.

tiguous United States, based on the ACCESS-CM2 model (which also contributed to CMIP6; Bi et al. 2013). Both datasets have been studied previously in the context of EEA in Kiriliouk and Naveau (2020) and Nanditha et al. (2024), respectively. The two applications are chosen to assess whether the sensitivity of attribution results to extremal dependence assumptions persists across both different precipitation regimes and different climate model structures, rather than reflecting only geographical variability within a single model or model-specific behaviour at a single location.

The spatial nature of both datasets means that the Stable Unit Treatment Value Assumption (SUTVA) does not hold. Spatial dependence among locations violates the assumption of unit-level independence, while spatial heterogeneity undermines the premise of a uniform treatment effect. Additionally, internal climate variability may mask or distort the true impact of the treatment, leading to potentially misleading inferences. These challenges are not unique to our study but are common to spatial applications of causal inference. To mitigate these violations in the context of event attribution, we first reduce heterogeneity by clustering locations with similar features. Clusters are introduced as overlapping local neighbourhoods around grid centroids to enforce approximate marginal homogeneity rather than as a global spatial partition. Within each cluster, where

spatial dependence may still be present, we fit the three statistical models introduced in Section 4.2, each designed to account for dependence structures in both the factual and counterfactual worlds. We then estimate and compare the attribution ratio (AR) across scenarios to assess the influence of human activity on precipitation extremes. This comparison relies on accurate estimation of extreme probabilities, as described in Section 4.2.6. To estimate these probabilities, we define a common event level corresponding to a high-impact threshold. In practice, this level is calibrated empirically at each location using the 90th percentile as a reference for the tail region (i.e., each component u_j of the vector $\boldsymbol{u}$ is set to the 90th percentile of the empirical distribution at location j, computed separately for the factual and counterfactual scenarios). Importantly, this calibration is used to identify comparable exceedance events in the factual and counterfactual worlds rather than to define different events in each world. The event of interest is therefore defined consistently across the factual and counterfactual scenarios, and the probability change is evaluated for the same class of extreme events, in line with the risk-based attribution framework (Otto, 2017). Estimation results for the probability of necessary causation (PN) across all statistical models and both applications are presented in Appendix A.5.

4.4.1 Precipitation in Europe using CNRM outputs

We analyse simulations from the Detection and Attribution Model Intercomparison Project (DAMIP) contribution to CMIP6 (Gillett et al., 2016), which provides coordinated experiments designed to separate the effects of anthropogenic and natural forcings for attribution studies. The CNRM dataset contains daily maximum precipitation values for the winter months (1st January 1985 to 31st December 2014; 2,707 time points) across 433 locations in central Europe. The factual and counterfactual scenarios are represented by DAMIP historical simulations, with the counterfactual world corresponding to runs forced by natural drivers only. To reduce potential nonstationarity over time, we follow Kiriliouk and Naveau (2020) and compute the 7-day maxima at each location, treating these as temporally stationary series within each world.

As outlined in Section 4.2, the three statistical models under consideration assume tail homogeneity across the spatial domain. To satisfy this assumption, we follow the approach of Castro-Camilo and Huser (2020) and partition the study region into homogeneous clusters, defined as regions where the marginal distributions of threshold exceedances are approximately stationary. Each model is then fitted locally within these clusters.

Algorithm 4.1 summarises the homogeneous-neighbourhood construction.

Algorithm 4.1: Construction of overlapping homogeneous local neighbourhoods

Input: Spatial observation locations, a regular grid of centroids, and an increasing sequence of candidate neighbourhood sizes $\mathcal{D} = \{D_1 < \cdots < D_L\}$, where D denotes the number of nearest observation locations included around a centroid.

Output: A collection of overlapping homogeneous local neighbourhoods.

1: **for** each centroid $\mathbf{s}_0$ **do**
2: Order all observation locations by their distance from $\mathbf{s}_0$.
3: For each $D \in \mathcal{D}$, form $\mathcal{N}_{\mathbf{s}_0,D}$ from the D observation locations nearest to $\mathbf{s}_0$; the nested candidate neighbourhoods therefore increase in size with D.
4: For each $D \in \mathcal{D}$, assess marginal homogeneity using the combined Hosking–Wallis and modified Anderson–Darling tests.
5: Set D_0 to the largest candidate size for which marginal homogeneity is not rejected.
6: Define $\mathcal{N}_{\mathbf{s}_0,D_0}$ as the local neighbourhood associated with $\mathbf{s}_0$.
7: **end for**
8: **return** the collection $\{\mathcal{N}_{\mathbf{s}_0,D_0}\}$, whose neighbourhoods may overlap and do not form a disjoint spatial partition.

Figure 4.4 displays three example neighbourhoods produced by Algorithm 4.1. For the three locations denoted in red (each taken from a different cluster), Figure 4.5 shows marginal qqplots for the three models introduced earlier. The top row corresponds to the counterfactual scenario, and the bottom row corresponds to the factual scenario. At the selected diagnostic locations, no model gives uniformly adequate marginal fits across both worlds; the HW fit is particularly poor in several displayed panels. To evaluate the performance of the dependence modeling, Figure 4.6 displays the bivariate conditional exceedance probability $\chi(u)$, as defined in Section 4.2.5, for each model. We focus on the three reference locations highlighted in red in Figure 4.5, pairing each with a neighbouring site from the same cluster (shown in green). These plots provide an empirical comparison

of the extremal behaviour implied by the fitted models, illustrating how their predictions diverge across levels of extremeness. At the selected pairs, the mGPD estimates are nearly constant in u and often lie above the empirical estimates near the upper tail, whereas the eFCM and HW curves are closer to the empirical curves in several panels. For all models, 95% pointwise confidence intervals are computed using nonparametric bootstrap. Taken together, Figures 4.5 and 4.6 show that the three models differ in both marginal and dependence fit. The attribution maps below are therefore compared as estimates obtained under alternative fitted tail specifications.

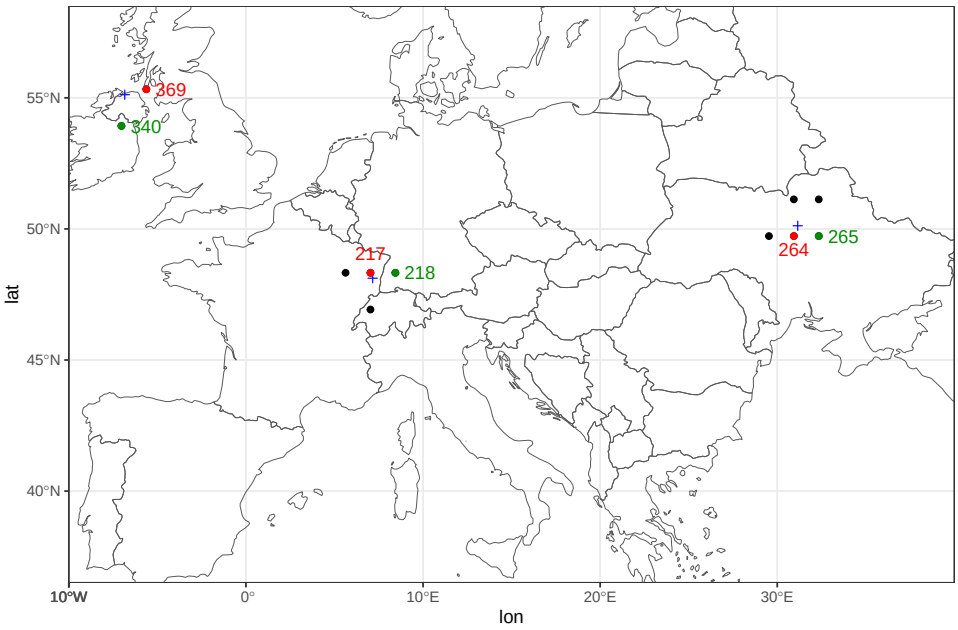

Figure 4.4: European CNRM study region and diagnostic locations. Blue crosses are the three grid centroids whose overlapping homogeneous neighbourhoods are constructed by Algorithm 4.1; black dots are observation sites in those neighbourhoods. Red sites are used for the marginal quantile diagnostics in Figure 4.5, and each green site is paired with the adjacent red site for the conditional-exceedance diagnostic in Figure 4.6.

Figure 4.7 displays maps of attribution ratios estimated under each of the three models, with v corresponding to the 5-year (left) and 50-year (right) return levels. Colors indicate the AR estimates, while transparency reflects uncertainty: more opaque colors correspond to narrower 90% confidence intervals and thus higher certainty, whereas greater transparency indicates increased uncertainty. For a clearer view of the spatial patterns in AR point estimates, excluding uncertainty, see Figure A.3 in Appendix A.4. To facilitate comparison, the eFCM and mGPD results are plotted using a common color scale spanning

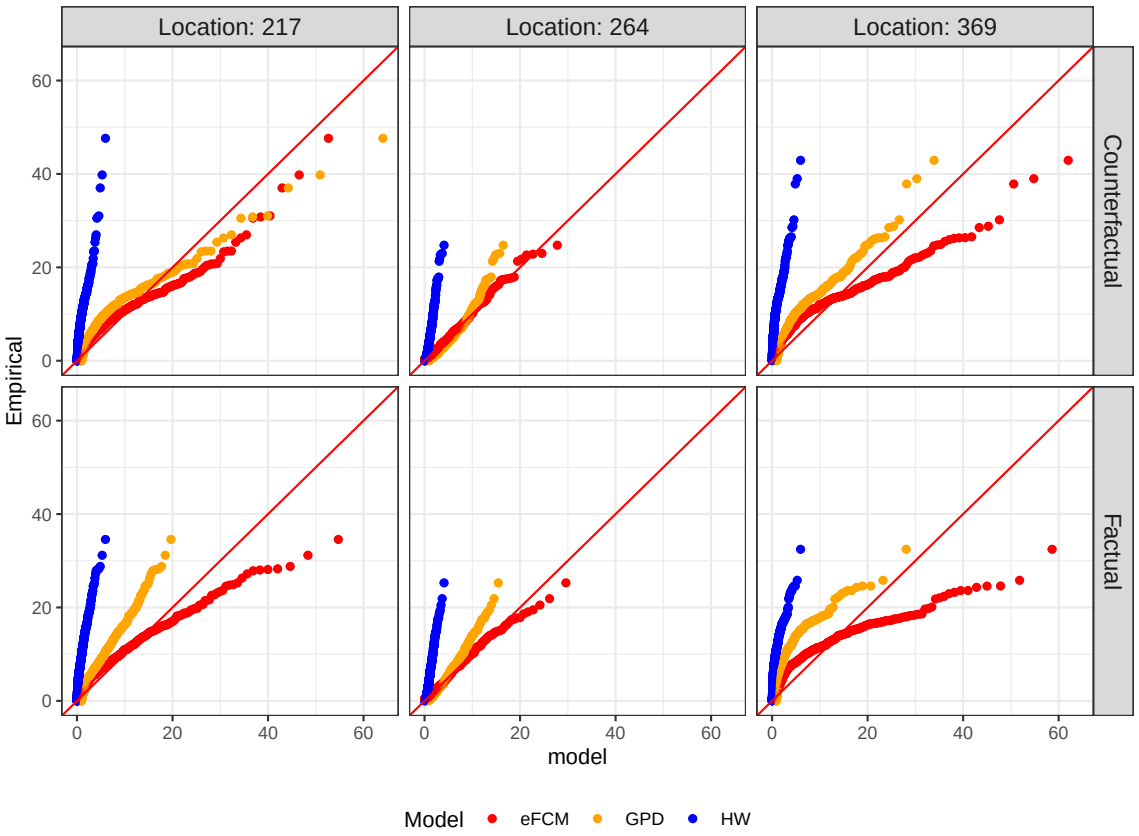

Figure 4.5: Marginal quantile–quantile diagnostics for the eFCM, mGPD, and HW fits at the three red locations in Figure 4.4. Each panel plots a fitted model quantile against the corresponding empirical precipitation quantile; agreement with the 45-degree line indicates adequate marginal fit. Columns identify locations, and rows show the counterfactual and factual climate-model worlds.

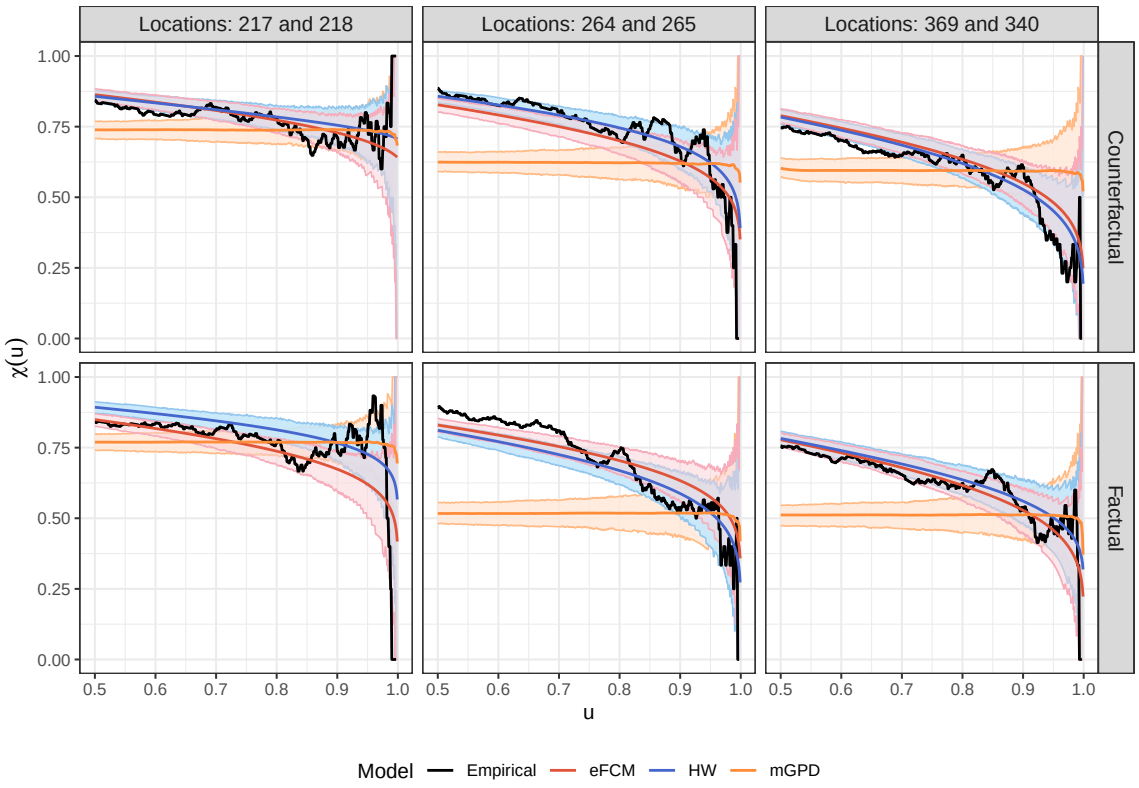

Figure 4.6: Conditional exceedance probabilities $\chi(u) = \Pr\{F_1(X_1) > u \mid F_2(X_2) > u\}$ for the selected European site pairs in Figure 4.4. The black line is the empirical estimator; red, blue, and orange are the eFCM, HW, and mGPD fits. Grey and model-coloured bands are 95% pointwise bootstrap confidence intervals. The columns identify site pairs and the rows distinguish the counterfactual and factual worlds.

$[-1, 1]$, while the HW model uses a narrower scale of $[-0.2, 0.2]$ to reflect the concentration of its estimates near zero. These spatial patterns should be interpreted in conjunction with the marginal and dependence diagnostics shown in Figures 4.5 and 4.6. In this context, the mGPD model, which imposes a constant tail dependence structure, produces smooth and interpretable spatial patterns but, in light of the results in Figure 4.6, may fail to capture localized variations in extremes. It also tends to yield wider confidence intervals in some regions, as evidenced by the lower opacity in its maps. In contrast, the eFCM and HW fits produce more spatially heterogeneous attribution maps. The eFCM intervals are narrower in many displayed regions. With respect to spatial attribution patterns, the HW and eFCM models show broadly similar structures. Both indicate a mix of weakly negative to neutral ARs in Spain and Portugal, with France showing more negative values in the south and near-zero signals in the north. Southern Scandinavia (including Denmark and southern Sweden) shows similarly subdued and consistent patterns across

both statistical models. Greater divergence appears in regions such as Italy, Germany, Southeast Europe, the UK, and Ireland, where the eFCM model reveals stronger negative ARs, in contrast to the mostly neutral or slightly positive estimates from the HW model. The eFCM maps exhibit less spatial coherence than the HW maps, a pattern consistent with empirical evidence that extremal dependence in precipitation weakens rapidly with distance.

Overall, the mGPD model suggests a decrease in risk attributable to human activity, whereas the eFCM indicates an increase, often displaying a spatial pattern similar to that of the HW model. Notably, regions where the mGPD detects an increase in risk, such as northern Italy and Switzerland in the 50-year return level map, are also identified by the eFCM, but are less consistently captured by the HW model.

4.4.2 Precipitation in the contiguous U.S. using ACCESS-CM2 outputs

Our second application focuses on precipitation extremes in the contiguous U.S. Here, we use simulations from the ACCESS-CM2 climate model, also participating in DAMIP (Gillett et al., 2016), under two scenarios: *hist-nat*, representing a counterfactual world with only natural forcings (solar radiation and volcanic aerosols), and *historical*, which includes both anthropogenic and natural forcings. The dataset comprises daily precipitation measurements at 251 observation locations spanning 12,418 days from January 1, 1981, to December 31, 2014. Consistent with our first application, we compute the 7-day maxima at each location. However, unlike the first application, we do not assume temporal stationarity, as we now consider data spanning the entire year. This introduces a seasonal effect that must be accounted for. To address this, we use a block bootstrap approach when estimating uncertainties.

We apply Algorithm 4.1 to identify spatially homogeneous regions, using a grid of 261 centroids spaced at 0.5-degree intervals in both latitude and longitude. This produces 261 overlapping local neighbourhoods, with the neighbourhood size D_0 selected separately at each centroid by the nested nearest-location homogeneity procedure specified in the algorithm. As in the first application, we evaluate marginal fits (Figure 4.8) and tail dependence behavior (Figure 4.9) at the locations highlighted by red and green dots in the

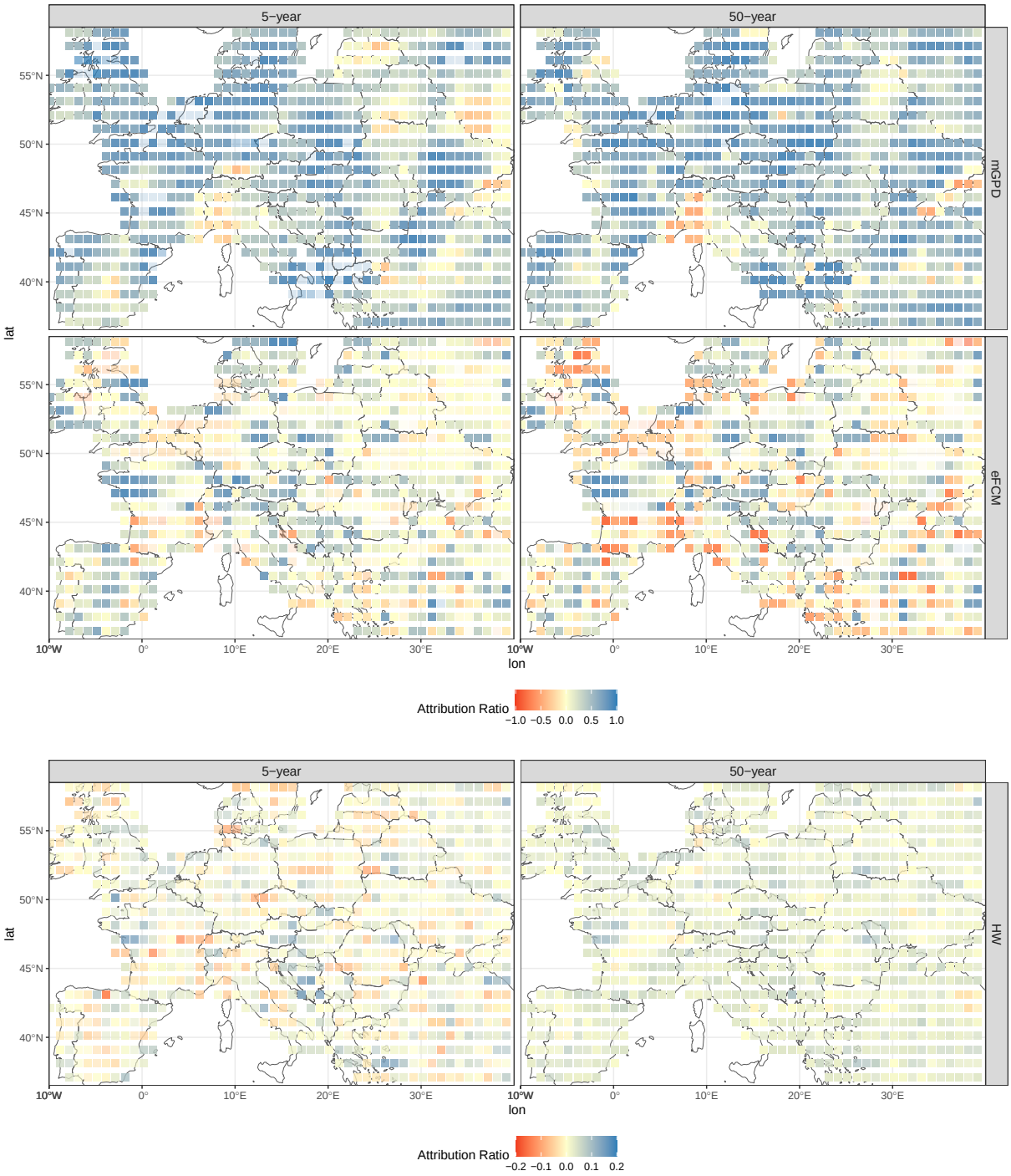

Figure 4.7: Spatial attribution ratios $\mathrm{AR} = (p_0 - p_1)/p_1$, where $p_t = \Pr\{\boldsymbol{w}^\top \boldsymbol{X}^{(t)} > v\}$ is defined in (4.1), for extreme precipitation across Europe. The left and right columns use 5-year and 50-year return levels for v. The mGPD and eFCM rows share a $[-1, 1]$ colour scale; the HW row uses its displayed narrower scale. Opacity represents the width of the 90% confidence interval, with more opaque points indicating narrower intervals. Negative AR means that the factual exceedance probability exceeds the counterfactual probability.

bottom left panel of Figure 4.10. For all models, 90% confidence intervals are computed using block bootstrap with three-month blocks to account for seasonal effects. At the

selected diagnostic locations, the eFCM and mGPD fits show closer marginal agreement than the HW fit, which underestimates upper-tail observations in the displayed panels. For the selected pairs in Figure 4.9, the eFCM and HW curves are generally closer to the empirical $\chi(u)$ curves than the mGPD curve.

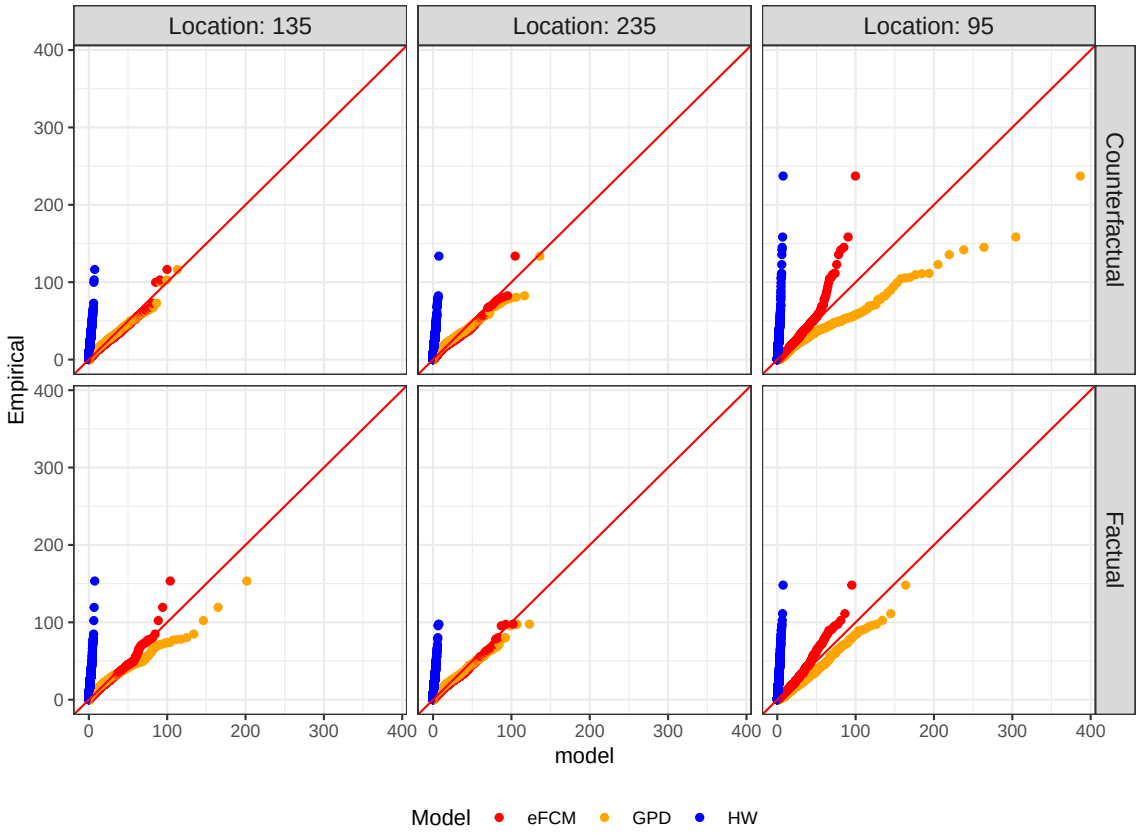

Figure 4.8: Marginal quantile–quantile diagnostics for the eFCM, mGPD, and HW fits at the selected U.S. locations marked in Figure 4.10. Each panel plots a fitted model quantile against the corresponding empirical precipitation quantile; agreement with the 45-degree line indicates adequate marginal fit. Columns identify locations, and rows show the counterfactual and factual climate-model worlds.

To facilitate the interpretation of the causal metrics, attribution ratios displayed in Figure 4.10 are classified into three categories: *negative* if they are below -0.05, *zero* if they are less than 0.05 in size, and *positive* if they are above 0.05. Confidence interval (CI) lengths are also grouped into bins of width 0.2 from 0 to 2, with a final category for lengths exceeding 2. In the visualizations, darker colors represent narrower CIs, indicating greater certainty. Recall that a negative AR implies that the probability of extreme precipitation is higher under the factual scenario (with anthropogenic forcing) than under the counterfactual, suggesting an increased risk due to human influence. The observed

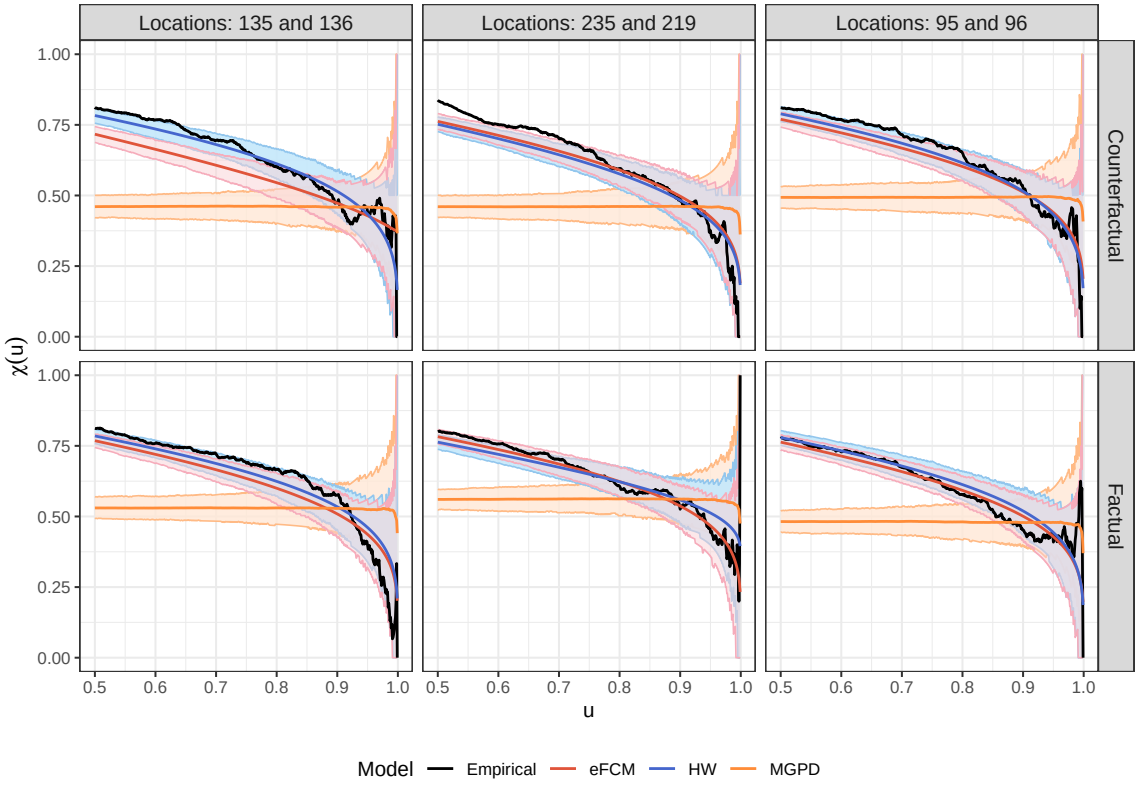

Figure 4.9: Conditional exceedance probabilities $\chi(u) = \Pr\{F_1(X_1) > u \mid F_2(X_2) > u\}$ for selected U.S. site pairs. The black line is the empirical estimator; red, blue, and orange are the eFCM, HW, and mGPD fits. Grey and model-coloured bands are 95% pointwise block-bootstrap confidence intervals. Columns identify site pairs and the rows distinguish the counterfactual and factual worlds.

spatial patterns, which should be interpreted together with the marginal and dependence diagnostics presented earlier in Figures 4.8 and 4.9, highlight clear differences in attribution patterns across statistical models. The mGPD tends to produce strongly polarized ARs, either highly positive or highly negative, and often suggests widespread positive ARs (blue). The mGPD combines a nearly constant fitted tail-dependence profile with a more polarised attribution map. In contrast, both the eFCM and HW models display spatially heterogeneous AR patterns, with mixed signals across regions and consistent visual encoding of uncertainty. However, key differences emerge: the HW model produces stronger and more polarized attribution signals, particularly at the 50-year return level, with many regions showing confidently positive or negative ARs. In contrast, the eFCM model yields more conservative estimates, with a greater prevalence of near-zero and uncertain values. Regional discrepancies are also notable; HW shows pronounced positive ARs in the Northeast and Midwest, which are less evident under the eFCM. Despite these differences, some broad regional trends are shared between statistical models, suggesting a degree of robustness in certain attribution signals while also highlighting sensitivity to statistical model choice and event rarity. Taken together, the diagnostics and attribution maps show that the inferred anthropogenic signal varies across the fitted tail models, especially at higher return levels. The eFCM results are qualitatively consistent with patterns reported by Nanditha et al. (2024) (their Supplementary Figure S6).

4.5 Discussion

This chapter examines how attribution metrics vary across alternative marginal and joint-tail specifications while holding the underlying climate simulations fixed within each application. The comparison covers the mGPD, eFCM, and HW models, two HW-based simulation studies, and two precipitation applications.

In Simulation Scenario 1, the selected dependence misspecification produces greater bias than the selected marginal distortion. In Simulation Scenario 2, the eFCM yields lower bias and RMSE than the mGPD across the HW settings considered. These results show that marginal specification, dependence specification, and model flexibility can materially affect estimated extreme-event probabilities.

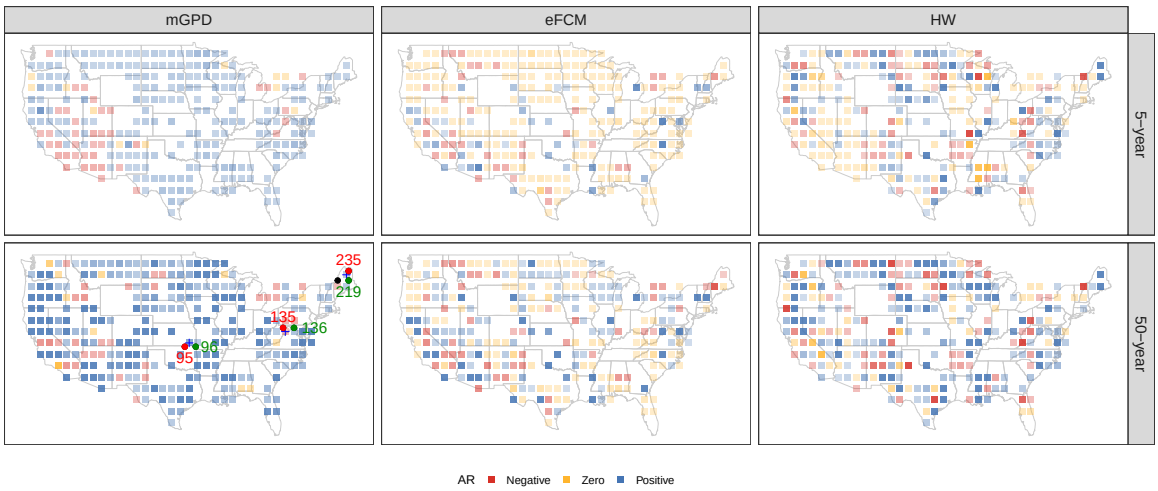

Figure 4.10: Spatial attribution ratios $\mathrm{AR} = (p_0 - p_1)/p_1$, with $p_t = \Pr\{\boldsymbol{w}^\top \boldsymbol{X}^{(t)} > v\}$ from (4.1), for extreme precipitation across the contiguous U.S. Rows use 5-year and 50-year return levels for v, and columns show the eFCM, HW, and mGPD fits. AR is classified as negative when below -0.05, near zero when between -0.05 and 0.05, and positive when above 0.05. Transparency encodes the width of the 90% confidence interval; lighter points have wider intervals.

The two applications also produce substantial differences in the magnitude and, in some regions, the sign of the estimated attribution ratios. The marginal and dependence diagnostics show that the three models capture different aspects of the observed tail behaviour. The attribution maps should therefore be read together with both sets of diagnostics.

Importantly, the homogeneous neighbourhood procedure used as a common modeling device across the three approaches identifies geographically compact regions by construction, whose statistical coherence is plausibly linked to physical drivers of extreme precipitation, such as orographic effects and regional circulation patterns. A formal analysis examining whether the resulting cluster geometries align with known physical precipitation regimes would be a valuable direction for future work.

Taken together, the simulations and applications show that tail-model choice is an important source of variation in extreme event attribution. In practice, comparing plausible tail specifications and reporting both marginal and dependence diagnostics provides a clearer account of the sensitivity of attribution conclusions.

While this analysis focuses on extreme precipitation, related sensitivity to tail specification may arise for other climate variables, including drought indices and compound events. Extending the comparison to a broader class of climate variables would clarify how the influence of marginal and dependence modelling changes across applications.

Beyond the climate attribution literature, these results are consistent with a long-standing body of work in decision analysis showing that assumptions about dependence structures can materially shape conclusions even when marginal distributions are well specified (Clemen and Winkler, 1985, 1999; Clemen et al., 2000). In that literature, it has been shown that adopting a single dependence model, such as independence or a convenient parametric form, is not a neutral choice, but an active modeling decision with substantive implications for inference and decision making. Related work has further demonstrated that, under partial probabilistic information, commonly used dependence assumptions may correspond to extreme points of the feasible set of joint distributions, potentially leading to overconfident or unrepresentative conclusions (Montiel and Bickel, 2012). Viewed through this lens, our results highlight an analogous issue in extreme event attribution, where attribution metrics are often evaluated under a specific joint tail model despite substantial uncertainty about extremal dependence. This broader perspective reinforces the importance of transparency and sensitivity analysis with respect to dependence assumptions in policy-relevant attribution studies.

Key messages

- Estimated attribution metrics vary materially across the fitted tail specifications in the simulations and precipitation applications.
- In the first HW simulation, the selected dependence misspecification produces greater bias than the selected marginal distortion.
- In the second HW simulation, the eFCM has lower bias and RMSE than the mGPD across the settings considered.
- The applications show model-dependent differences in the magnitude and spatial pattern of the estimated attribution ratios.
- Marginal and dependence diagnostics should be assessed jointly when interpreting attribution results.

Chapter 5

Extreme quantile treatment effects

This chapter develops a tail-calibrated estimating-equation framework for causal effects at extreme quantiles. By integrating causal adjustment with extreme-value tail modelling, the proposed approach stabilises inference in regimes where the target quantile lies so far into the tail that standard quantile methods become unreliable.

A version of this chapter has been submitted for publication under the title *Tail-Calibrated Estimation of Extreme Quantile Treatment Effects*. Preprint available at arXiv:2603.23309.

5.1 Introduction

Estimating causal effects at extreme quantiles is central in applications where rare, high-impact outcomes drive risk and decision-making. Examples arise in climate attribution, financial losses, environmental hazards and health extremes, where the interest lies in changes in the severity of the most extreme events rather than in average effects. A natural framework for studying distributional causal effects is that of quantile treatment effects (QTEs), widely explored in classical statistics and econometrics (Doksum, 1974; Lehmann and D'Abrera, 1975). Important examples include that of Firpo (2007), who developed semiparametric estimation of QTEs under unconfoundedness, and (Ma et al., 2023), who extended Firpo's work to more complex data structures and missingness. Zhang (2018a) formalised the conditions under which these types of estimators retain their desired asymptotic properties by distinguishing between intermediate, moderately extreme and extreme quantile regimes according to the effective tail sample size. Zhang (2018a) concluded that asymptotic normality can be recovered for intermediate quantiles, but inference becomes ill-posed as the target quantile moves further into the tail. Formal causal analysis of tail quantiles remains, therefore, methodologically challenging.

One potential approach to causal inference at extreme quantile levels is to build upon advancements in quantile regression for extreme quantiles, as developed by Chernozhukov (2005), Chernozhukov and Fernández-Val (2011) and summarised in Chernozhukov et al. (2018). However, these methods do not explicitly incorporate a causal mechanism. As a result, adaptations to a causal framework are usually limited to settings where controlling for confounders is unnecessary, and thus cannot be applied to observational data. Recently, Cheng and Li (2025) proposed a method for causal quantile estimation via a general inverse estimating equation framework. Their approach provides a unifying perspective on quantile identification, but their theory primarily applies to interior quantiles.

Recent work has turned to extreme value theory (EVT) to facilitate estimation and inference for extreme quantiles beyond the range of observed values. The classical reference by De Haan and Ferreira (2007) provides the theoretical foundation for tail extrapolation under regular variation assumptions. For heavy-tailed distributions, the tail quantile function follows a power-law representation governed by the extreme value index (EVI), which is typically estimated using classical procedures such as the Hill estimator (Hill, 1975). Within this framework, Wang et al. (2012) and Xu et al. (2022) extended EVT-based methods to conditional and regression settings, while Deuber et al. (2024) were among the first to explicitly link EVT with causal inference for quantile treatment effects (QTEs) in observational studies. Although this work constitutes a significant advancement, the approach relies on heavy-tailed (Fréchet-domain) assumptions and employs a two-step estimation procedure followed by extrapolation. Alternative approaches to QTE at extreme levels have emerged from specific applied settings, such as that of Bhuyan et al. (2023), who combine tools of causal inference and EVT to study London traffic flow data.

Taken together, the existing literature reveals a significant conceptual gap. Causal identification and extreme value asymptotics have largely been developed as separate domains, and current methodologies typically address confounding adjustment and tail extrapolation in a sequential manner rather than integrating them within a single inferential framework. As a result, there is no unified approach that provides formally identified, statistically valid inference for causal effects at extreme quantile levels. With this in mind, this chapter introduces the Tail-Calibrated Inverse Estimating Equation (TIEE) framework, which provides a new inferential paradigm for assessing causal effects in the extreme

tails. Instead of focusing on a single extreme quantile, TIEE aggregates information across a continuum of tail quantile levels using an estimating equation approach specifically calibrated for extremes. This methodology incorporates causal identification directly into the estimation process, resulting in a well-defined inferential object in settings where conventional quantile-based methods are inadequate, and unifies causal signal functions with extreme value theory within a single coherent framework.

Our contributions are fourfold. First, we introduce a general class of tail-calibrated estimating equations that provides a well-posed identification strategy for extreme quantile treatment effects without requiring restrictive tail assumptions, such as limiting analysis to the Fréchet domain. Second, we develop the corresponding estimator, implement computationally stable procedures, and demonstrate how EVT components inform the causal signal function to facilitate extrapolation beyond observed extremes. Third, we establish theoretical properties of the estimator, including identification, consistency, and asymptotic normality, and provide variance expressions that account for uncertainty arising from both causal adjustment and tail modelling. Finally, through extensive simulations across light- and heavy-tailed regimes, we show substantial reductions in bias and variance compared to existing extreme-QTE methods, and we illustrate the approach in an observational study of extreme precipitation in the Austrian Alps.

The remainder of the chapter is organised as follows. Section 5.2 provides background on causal inference, extreme quantile regimes, and the inverse estimating equation framework. Section 5.3 introduces the TIEE framework and its integration with causal signal functions and EVT-based tail modelling. Section 5.4 presents the main theoretical results and inference procedures for the extreme quantile treatment effect. Section 5.5 reports simulation results under heavy- and light-tailed designs, including sensitivity and robustness analyses. Section 5.6 applies the method to Austrian extreme precipitation data, and Section 5.7 concludes. The code to reproduce the simulation studies in Section 5.5, and the data and code to reproduce the data application in Section 5.6 are freely available at `https://github.com/MengranLi-git/TIEE`.

A note on notation. Throughout this chapter, $D \in \{0,1\}$ denotes a binary treatment indicator, with $D = 0$ corresponding to no intervention and $D = 1$ to intervention. We use $d \in \{0,1\}$ to denote generic values of D, and $Y(d)$ to denote the corresponding potential outcome, with the subscript omitted in purely descriptive contexts. The vector $\boldsymbol{X}$ denotes observed covariates, and the data consist of independent copies $W(i) = (Y(i), D(i), \boldsymbol{X}(i))$, $i = 1, \ldots, n$, of $W = (Y, D, \boldsymbol{X})$. We write $F_{Y,d}(\cdot)$ for the marginal distribution of Y under treatment level d Additionally, we distinguish between quantile levels and quantile values. Quantile levels are denoted by τ or $p \in (0,1)$, depending on context, and are treated as fixed when defining the estimand. $Q_{Y(d)}(\tau)$ denotes the marginal quantile value associated to the quantile level τ under treatment d. To avoid notation clutter when needed and highlight the estimation target, we write $\theta_{d,\tau} = Q_{Y(d)}(\tau)$. When no confusion arises, we also use q or $q(p)$ to denote a generic quantile function evaluated at level p.

5.2 Background

Building on the causal inference foundations introduced in Section 2.1, this section defines the extreme quantile treatment effect and reviews the main ingredients required for its estimation. Let $Y_i(1) - Y_i(0)$ denote the causal effect of the treatment for observation i. To identify and estimate treatment effects from observational data, we require the following standard assumptions.

Assumption 5.1 (Identifiability conditions).

1. Consistency: The potential outcome is consistent with the observed outcome.
2. Stable Unit Treatment Value Assumption (SUTVA): The potential outcomes for one unit are not affected by the treatments assigned to other units, and the treatment levels are the same for all units.
3. Unconfoundedness: $(Y(0), Y(1)) \perp D|\boldsymbol{X}$, where $\boldsymbol{X}$ is a set of covariates or confounders and the treatment D is completely randomised within strata defined by $\boldsymbol{X}$. This means that the outcomes are independent of the treatment given the set of confounders.
4. Overlap: $0 < \Pr(D = 1 \mid \boldsymbol{X}) < 1$, for all $\boldsymbol{X}$ in the support, which means that both treatment states must be possible for every covariate pattern.

Whilst the average treatment effect (ATE), defined as $\mathbb{E}\{Y(1)-Y(0)\}$, remains the most commonly reported causal estimand, it only characterises the location shift between treatment distributions. The quantile treatment effect (QTE) provides a more comprehensive characterisation of distributional differences and is defined as

$$\delta(\tau) = Q_{Y(1)}(\tau) - Q_{Y(0)}(\tau) = \theta_{1,\tau} - \theta_{0,\tau} \quad \tau \in [0,1], \tag{5.1}$$

where $\theta_{d,\tau} = Q_{Y(d)}(\tau) = \inf\{y : F_{Y,d}(y) \geq \tau\}$ is the τ-quantile of the potential outcome $Y(d)$.

5.2.1 Existing approaches to quantile treatment effect estimation

Section 3.3 situates QTE methods within the broader problem of causal inference for extremes. This subsection returns to the specific estimating strategies and distributional representations that motivate the construction of TIEE.

Without imposing distributional assumptions, Firpo (2007) proposed a QTE estimator for $\tau \in (0,1)$ obtained by reweighting the quantile loss function as

$$\widehat{Q}_{Y(d)}(\tau) = \arg\min_{q \in \mathbb{R}} \sum_{i=1}^{n} w_{d,i} \rho_\tau(Y_i(d) - q),$$

where $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$ is the check function and $w_{d,i}$ are inverse probability weights (IPWs) constructed from the propensity score, defined as the conditional probability of receiving treatment given the observed covariates.

An alternative approach estimates QTE via conditional distributions, as in Zhang et al. (2012). Let $G_{Y(d)}(\cdot \mid \boldsymbol{x}, \boldsymbol{\nu})$ be a parametric model for $F_{Y_d}(\cdot \mid \boldsymbol{x})$, where G is a known function and $\boldsymbol{\nu}$ is an unknown, finite-dimensional parameter. The marginal distribution of Y under treatment d can be estimated as

$$\widehat{F}_{Y(d)}(y) = \widehat{\mathbb{E}}_{\boldsymbol{X}}\{\Pr(Y(d) \leq y \mid \boldsymbol{X})\} = \frac{1}{n} \sum_{i=1}^{n} G_{Y(d)}(y \mid \boldsymbol{X}(i); \hat{\boldsymbol{\nu}}),$$

where $\widehat{\mathbb{E}}_{\boldsymbol{X}}(\cdot)$ denotes expectation with respect to the marginal distribution of $\boldsymbol{X}$ and $\widehat{\boldsymbol{\nu}}$ is an estimate of $\boldsymbol{\nu}$. The above can be inverted to obtain an estimate of the desired quantile. Zhang et al. (2012) adopts this strategy under the assumption that $G(\cdot)$ is a Gaussian distribution after a Box-Cox transformation of Y.

Ma et al. (2023) proposed estimating QTEs using a link between marginal quantiles and quantile regression. Specifically, assuming the linear quantile regression model

$$\widehat{Q}_{Y(d)}(\tau \mid \boldsymbol{X}) = \boldsymbol{X}^T \boldsymbol{\beta}(\tau), \quad 0 < \tau < 1.$$

Let $M(Y, \theta_{d,\tau}) = \mathbb{1}\{Y \leq \theta_{d,\tau}\} - q$. The link between marginal and conditional quantiles is constructed using

$$\mathbb{E}_Y[M(Y, \theta_{d,\tau})] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}_{Y|\boldsymbol{X}}(M(Y, \theta_{d,\tau}) \mid \boldsymbol{X})] = 0,$$

where the conditional expectation can be expressed as an integral over the conditional quantile level τ, i.e.,

$$\mathbb{E}_{Y|\boldsymbol{X}}(M(Y, \theta_{d,\tau}) \mid \boldsymbol{X}) = \int_0^1 M(F_Y^{-1}(q \mid \boldsymbol{X}), \theta_{d,\tau}) \, \mathrm{d}q.$$

Ma et al. (2023) estimate the conditional expectation using the quantile regression model

$$\widehat{\mathbb{E}}_{Y|\boldsymbol{X}}(M(Y, \theta_{d,\tau}) \mid \boldsymbol{X}) = \int_0^1 M(\boldsymbol{X}^T \boldsymbol{\beta}(q), \theta_{d,\tau}) \, \mathrm{d}q, \tag{5.2}$$

where $\boldsymbol{\beta}(q)$ is estimated by minimising the usual quantile regression loss function. Thus, $\theta_{d,\tau}$ is the unique value solving

$$\widehat{\mathbb{E}}_{\boldsymbol{X}} \left[\int_0^1 M(\boldsymbol{X}^T \boldsymbol{\beta}(q), \theta_{d,\tau}) \, \mathrm{d}q \right] = 0. \tag{5.3}$$

This equation defines the marginal estimating equation used to recover $\theta_{d,\tau}$. In practice, (5.3) is solved by replacing $\widehat{\mathbb{E}}_{\boldsymbol{X}}(\cdot)$ by the sample average, with the integral evaluated using a trimmed domain to avoid instability at extreme quantile levels. This approach yields estimators with smaller standard errors than Firpo (2007).

5.2.2 Extreme quantile treatment effects

The tail approximation used below builds on the EVT framework in Section 2.2, including the GEV limit in Eq. (2.18) and the role of the shape parameter. The following discussion connects this background to the extreme-QTE regimes reviewed in Section 3.3.

Whilst QTE estimators provide a comprehensive overview of treatment effect distributions, standard asymptotic properties no longer hold when τ approaches 0 or 1. Considering the lower tail and following Zhang (2018a), we distinguish three asymptotic regimes for extreme quantile levels $\tau_n \to 0$, defined by how fast the effective sample size $n\tau_n$ shrinks:

$$\begin{aligned} \text{Intermediate:} \quad & \tau_n \to 0 \text{ and } n\tau_n \to \infty, \\ \text{Moderately extreme:} \quad & \tau_n \to 0 \text{ and } n\tau_n \to d > 0, \\ \text{Extreme:} \quad & \tau_n \to 0 \text{ and } n\tau_n \to 0. \end{aligned}$$

The first regime still admits Gaussian limits for suitably normalised estimators, whereas in the latter two regimes, empirical quantiles become unstable and classical root-n theory fails. Zhang (2018a) established that Firpo's (2007) estimator is asymptotically normal for intermediate quantiles and developed b-out-of-n bootstrap procedures for moderately extreme cases. However, empirical quantile-based methods become unreliable when $n\tau_n \to 0$, as few or no observations fall in the tail region. In the supplementary materials (Zhang, 2018b), the author also proposes a Pickands-type estimator for the extreme value indices (EVIs) of the potential outcome distributions, adapted to the causal inference setting in which potential outcomes are only partially observed. The EVI quantifies the heaviness or lightness of the tail of a distribution. Specifically, in EVT, one typically assumes an i.i.d. sample $Y(1) \ldots, Y(n)$ from distribution G that lies in the max-domain of attraction of a generalised extreme value distribution, meaning there exist sequences of constants $\{a_n \geq 0\}$ and $\{b_n\}$ such that

$$\lim_{n\to\infty} G^n(a_n y + b_n) = \exp\left\{-(1+\gamma y)^{-1/\gamma}\right\}, \tag{5.4}$$

where γ is the EVI. Heavy-tailed distributions correspond to $\gamma > 0$, light-tailed to $\gamma = 0$, and bounded tails to $\gamma < 0$. In Zhang (2018a), if $\hat{q}_d(\tau)$ denote the estimator of the τ-th quantile of $Y(d)$, the EVI γ_d is estimated as

$$\hat{\gamma}_d = -\sum_{r=1}^{R} w_r \frac{\log(\hat{q}_d(ml^r\tau_n) - \hat{q}_d(l^r\tau_n)) - \log\big(\hat{q}_d(ml^{r-1}\tau_n) - \hat{q}_d(l^{r-1}\tau_n)\big)}{\log l}, \tag{5.5}$$

where $l > 0$ is a spacing parameter, $m > 1$, $\{w_r\}_{r=1}^{R}$ are weights satisfying $\sum_{r=1}^{R} w_r = 1$, and τ_n is an intermediate quantile index.

To enable estimation beyond the range of observed data, Deuber et al. (2024) integrated extreme value theory with causal inference. Specifically, they proposed the following causal Hill estimators for the EVI:

$$\widehat{\gamma}_0^H := \frac{1}{n\tau_n}\sum_{i=1}^{n}\left(\log(Y(i)) - \log(\widehat{q}_0(1-\tau_n))\right)\frac{1-D(i)}{1-\widehat{\pi}(\boldsymbol{X}_i)}\mathbb{1}_{Y(i)>\widehat{q}_0(1-\tau_n)},$$
$$\widehat{\gamma}_1^H := \frac{1}{n\tau_n}\sum_{i=1}^{n}\left(\log(Y(i)) - \log(\widehat{q}_1(1-\tau_n))\right)\frac{D(i)}{\widehat{\pi}(\boldsymbol{X}_i)}\mathbb{1}_{Y(i)>\widehat{q}_1(1-\tau_n)},$$

where $\widehat{\pi}(\boldsymbol{X}_i)$ is the estimated propensity score, estimated via a logistic sieve estimator, with standard logistic regression used as a low-dimensional special case. They then extrapolate Firpo's intermediate quantile estimates to extreme levels as

$$\widehat{q}_d(1-p_n) = \widehat{q}_d(1-\tau_n)\left(\frac{\tau_n}{p_n}\right)^{\widehat{\gamma}_d^H}, \tag{5.6}$$

where p_n satisfies $p_n \to 0$ and $np_n \to d \geq 0$ (moderately extreme or extreme cases), whilst τ_n satisfies $\tau_n \to 0$ and $n\tau_n \to \infty$ (intermediate case). The relationship in (5.6) is approximated and applied only to heavy-tailed distributions ($\gamma > 0$), which means the Hill estimators handle quantiles beyond the data range but are restricted to heavy-tailed distributions.

Bhuyan et al. (2023) proposed a semi-parametric approach to extreme QTE estimation that reconstructs the entire counterfactual outcome distributions rather than extrapolating empirical quantiles directly. Their method targets the marginal distribution by modelling the conditional quantile function with a bulk-tail decomposition using a data-driven transition level τ_u. Extreme marginal quantiles are then obtained by integrating over co-

variates and inverting the estimated marginal distribution. This approach reconstructs the marginal tail distribution by modelling the entire conditional outcome distribution, requiring correct specification of the conditional quantile process and repeated threshold modelling across covariates.

5.2.3 The inverse estimating equation framework

The inverse estimating equation (IEE) framework introduced by Cheng and Li (2025) provides a general strategy for estimating quantiles of potential outcome distributions by exploiting the defining relationship between quantiles and threshold-transformed means. To streamline notation, we present the framework in the marginal setting, without conditioning on a subpopulation index. The formulation in Cheng and Li (2025), which allows all quantities to be defined conditionally on an index V, is recovered as a special case. For any threshold $\theta \in \mathbb{R}$, the mean of the threshold-transformed potential outcome is

$$\tau_d(\theta) = \mathbb{E}[\mathbb{1}\{Y(d) \leq \theta\}] = \Pr(Y(d) \leq \theta), \tag{5.7}$$

which corresponds to the distribution function of $Y(d)$ evaluated at θ. The object $\tau_d(\theta)$ plays a central role in the IEE framework, as quantiles are defined as its inverse. Under Assumption 5.1 and the additional requirement of point identification in Cheng and Li (2025), $\tau_d(\theta)$ can be identified from observed data through a signal function $s_d(W, \theta, \zeta)$ satisfying

$$\mathbb{E}[s_d(W, \theta, \zeta)] = \tau_d(\theta), \quad \text{for all } \theta \in \mathbb{R}, \tag{5.8}$$

where $W = (Y, D, \boldsymbol{X})$ denotes the observed data and ζ collects nuisance parameters. The signal function s is constructed so that its expectation recovers the threshold-transformed mean, without requiring a parametric model for the outcome distribution. In practice, estimation is carried out using a zero-mean estimating equation. To this end, define the estimating function

$$g_d(W, \tau, \theta, \zeta) = s_d(W, \theta, \zeta) - \tau,$$

where $\tau \in [0, 1]$ and ζ collects all nuisance components. By construction,

$$\mathbb{E}[g_d(W, \tau, \theta, \zeta)] = 0 \quad \Longleftrightarrow \quad \tau = \tau_d(\theta).$$

This representation shows that the zero-mean estimating equation commonly used in the IEE framework is algebraically equivalent to identifying $\tau_d(\theta)$ through the signal function in (5.8). The two formulations differ only in whether the probability level τ is placed on the left- or right-hand side of the equation.

Quantile identification follows by inversion. Let $\theta^\star_{d,q} = Q_{Y(d)}(q)$ denote the q-quantile of $Y(d)$, for some $q \in (0,1)$. By definition, $\tau_d(\theta^\star_{d,q}) = q$, and hence $\theta^\star_{d,q}$ satisfies the inverse estimating equation

$$\mathbb{E}\Big[g_d(W, q, \theta^\star_{d,q}, \zeta)\Big] = 0. \tag{5.9}$$

Viewed as a function of θ for fixed q, equation (5.9) characterises the quantile $\theta^\star_{d,q}$ as a root of the estimating equation. Cheng and Li (2025) show that, under suitable regularity conditions, this root is unique. In practice, $\theta^\star_{d,q}$ is estimated by solving the empirical analogue of (5.9), replacing ζ' by an estimate $\widehat{\zeta'}$.

When the target quantile level is bounded away from 0 and 1, and the density of $Y(d)$ is well behaved in a neighbourhood of $\theta^\star_{d,\tau}$, the empirical version of the estimating equation (5.9) admits a stable solution and standard asymptotic theory applies. As the target quantile level approaches the extremes, however, the behaviour of the IEE changes fundamentally. Empirical information in the neighbourhood of $\theta_{0,\tau}$ becomes increasingly sparse, the conditional density may be very small or vanishing, and small perturbations in the estimating equation can result in large changes in the solution. Consequently, the inversion implicit in (5.9) becomes ill-conditioned, leading to numerical instability and a breakdown of classical asymptotic arguments. These difficulties are not specific to a particular choice of estimating function or causal signal, but reflect a structural limitation of pointwise quantile inversion when extrapolation beyond the observed support is required. Overcoming this limitation motivates a reformulation of the estimating equation itself, which is the focus of the next section.

5.3 Tail-calibrated inverse estimating equation framework

This section introduces the Tail-Calibrated Inverse Estimating Equation (TIEE) framework for estimating extreme quantile treatment effects. The main contribution is a reformulation of the inverse estimating equation approach that replaces pointwise inversion at a fixed quantile level with a single integrated moment condition. This construction fundamentally alters the nature of the estimation problem. Indeed, by aggregating information across quantile levels, the resulting estimating equation remains well defined in the extreme regime. We begin by introducing the regularity conditions required to support inference.

Assumption 5.2 (Regularity conditions on potential outcomes). For each treatment level $d \in \{0, 1\}$, the following conditions hold:

1. The potential outcome $Y(d)$ is continuously distributed with conditional density $f_d(\cdot)$.
2. There exists a threshold y_0 such that the density $f_d(y)$ is monotone in the upper tail for all $y > y_0$.
3. For any fixed compact set C contained in the support of $Y(d)$, the density is bounded away from zero, that is, $\inf_{y \in C} f_d(y) > 0$.
4. The distribution $F_d(\cdot)$ belongs to the maximum domain of attraction of an extreme value distribution with extreme value index ξ_d (see (5.4)).

Assumption 5.2 adapts standard regularity conditions used in the quantile treatment effect literature (Firpo, 2007; Cheng and Li, 2025) to settings where inference targets the extreme tail. In particular, it relaxes the classical requirement that the outcome density be uniformly bounded away from zero, which is incompatible with unbounded outcomes. In particular, Condition 5.2(4) assumes that the potential outcome distributions belong to a maximum domain of attraction, providing a semi-parametric characterisation of tail behaviour that supports consistent tail approximation using extreme value methods. This assumption underpins the tail calibration step of the TIEE framework and does not restrict inference to a specific parametric tail model.

5.3.1 From inverse estimating equations to tail calibration

The classical inverse estimating equation (IEE) framework identifies quantiles by exploiting the pointwise relationship between a threshold and its associated distribution function. Specifically, for each treatment level d, the mapping

$$\theta \mapsto \tau_d(\theta) = \mathbb{E}[\mathbb{1}\{Y(d) \leq \theta\}]$$

is strictly increasing, and the target quantile $\theta_{d,\tau}$ is defined as the unique solution to $\tau_d(\theta) = \tau$. Estimation proceeds by constructing a zero-mean estimating equation for $\tau_d(\theta)$ at a fixed probability level τ and inverting it pointwise. This approach relies critically on local regularity. Specifically, the distribution function must vary sufficiently fast in a neighbourhood of $\theta_{d,\tau}$ and there must be adequate empirical information near the target quantile. In extreme regimes, these conditions fail. The effective sample size in the tail collapses, the distribution function becomes nearly flat, and the pointwise inversion implicit in the IEE framework becomes ill-conditioned. As a result, even well-identified estimating equations can lead to unstable or undefined quantile estimates.

The Tail-Calibrated Inverse Estimating Equation (TIEE) framework addresses this limitation by reformulating the object that is inverted. Rather than treating the pointwise threshold mean $\tau_d(\theta)$ as the primitive object, TIEE starts from a distribution-level representation that aggregates information across quantile levels. Specifically, note that for any $\theta \in \mathbb{R}$, we can write (5.7) as

$$\tau_d(\theta) = \int_0^1 \mathbb{1}\{Q_{Y(d)}(p) \leq \theta\}\, dp.$$

This identity expresses the distribution function as the measure of quantile levels whose values lie below θ. This identity holds independently of local density behaviour and does not rely on pointwise inversion.

Motivated by this representation, we define the random integrated signal

$$S_d(W, \theta) := \int_0^1 s_{\mathrm{TIEE},d}(W, \theta, \zeta, p)\, dp, \tag{5.10}$$

and its population counterpart

$$S_d(\theta) := \mathbb{E}[S_d(W,\theta)] = \int_0^1 \mathbb{E}[s_{\text{TIEE},d}(W,\theta,\zeta,p)]\,dp, \tag{5.11}$$

where ζ denotes nuisance parameters and the last equality in (5.11) is true under mild integrability conditions (e.g., $\mathbb{E}[\int_0^1 |s_{\text{TIEE},d}(W,\theta,\zeta,p)|\,dp] < \infty$). The identifying requirement of the TIEE framework is the *integrated unbiasedness condition*

$$S_d(\theta) = \int_0^1 \mathbb{E}[s_{\text{TIEE},d}(W,\theta,\zeta,p)]\,dp = \tau_d(\theta) \quad \text{for all } \theta \in \mathbb{R}. \tag{5.12}$$

Under the condition in (5.12), identification is now achieved through a global probability balance rather than a local condition at a single quantile level. The extreme quantile $\theta_{d,\tau}$ is therefore characterised as the unique solution to the integrated estimating equation

$$\begin{aligned} S_d(\theta) = \tau \quad &\text{or equivalently} \quad \mathbb{E}[g_{\text{TIEE},d}(W,\theta,\tau,\zeta)] = 0, \\ &\text{with} \quad g_{\text{TIEE},d}(W,\theta,\tau,\zeta) = S_d(W,\theta) - \tau. \end{aligned} \tag{5.13}$$

This reformulation preserves the core logic of the IEE framework (quantiles are still obtained as roots of estimating equations) while stabilising the inversion by aggregating information across quantile levels. As a result, the estimating equation remains well-defined even when empirical information near the target quantile is sparse. The integrated formulation naturally induces a decomposition into a data-rich body region and a tail region that requires extrapolation. Indeed, we can express (5.10) as

$$S_d(W,\theta) = \underbrace{\int_0^u s_{\text{TIEE},d}(W,\theta,\zeta,p)\,\mathrm{d}p}_{\text{body}} + \underbrace{\int_u^1 s_{\text{TIEE},d}(W,\theta,\zeta,p)\,\mathrm{d}p}_{\text{tail}},$$

where $u \in (0,1)$. We can then express the condition in (5.13) as

$$\mathbb{E}\big[g_{\text{TIEE},d}(W,\theta,\tau,\zeta)\big] = \mathbb{E}\left[\int_0^1 \big\{s_{\text{TIEE},d}(W,\theta_{d,\tau},\zeta,p) - \tau_{\text{eff}}\big\}\,\mathrm{d}p\right] = 0,$$

where $\tau_{\text{eff}} = \frac{\tau - p_u}{1-p_u} = \Pr(Y_d \le \theta_{d,\tau} \mid Y(d) > u)$ denotes the quantile level relative to the tail distribution.

5.3.2 Causal signal functions

The integrated estimating equation introduced in Section 5.3.1 is deliberately agnostic about how the distribution of the potential outcome $Y(d)$ is identified from observed data. Causal identification enters the TIEE framework solely through the choice of the signal function embedded in the estimating equation, while the structure of the estimator itself remains unchanged. This separation decouples identification from estimation and allows the proposed method to accommodate a wide range of causal designs within a unified framework.

When the unbiasedness condition in (5.12) holds, the signal function s_{TIEE} provides an unbiased representation of the marginal distribution of $Y(d)$. In particular, this condition ensures that, after integration over p, the signal recovers the marginal distribution function of the potential outcome, while the integrated estimating equation in (5.13) determines how the corresponding quantile is obtained. Importantly, the signal function is indexed by the quantile level p, even though the target estimand is a single extreme quantile $\theta_{d,\tau}$. This indexing is essential: TIEE does not attempt to identify the extreme quantile directly, but instead reconstructs the entire distribution function through aggregation across quantile levels, thereby avoiding ill-conditioned pointwise inversion. The flexibility of the framework follows from the fact that (5.12) can be satisfied by a variety of signal constructions, each corresponding to a different set of identifying assumptions. We briefly describe two canonical examples that are relevant for the applications considered in this chapter.

Naïve signal under randomisation. In a randomised experiment, treatment assignment is independent of the potential outcomes. In this setting, a valid signal function is given by the indicator itself,

$$s_{\mathrm{TIEE}}(W, \theta, p) \;=\; \mathbb{1}\{Q_{Y(d)}(p) \le \theta\}. \tag{5.14}$$

This signal trivially satisfies the unbiasedness condition in (5.12), but it is not valid in observational studies where treatment assignment depends on covariates.

Inverse probability weighted signal. In observational settings, causal identification can be achieved under unconfoundedness and overlap assumptions through inverse probability weighting. Let $\pi_d(\boldsymbol{X};\zeta) = \Pr(D = d \mid \boldsymbol{X})$ denote the propensity score. A valid TIEE signal is then given by

$$s_{\text{TIEE}}(W,\theta,\zeta,p) \;=\; \frac{\mathbb{1}\{D=d\}}{\pi_d(\boldsymbol{X};\zeta)}\,\mathbb{1}\{Q_{Y(d)}(p)\le\theta\}.$$

Under standard causal assumptions, this signal satisfies (5.12), since

$$\begin{aligned}\int_0^1 \mathbb{E}\left[\frac{\mathbb{1}\{D=d\}}{\pi_d(\boldsymbol{X};\zeta)}\,\mathbb{1}\{Q_{Y(d)}(p)\le\theta\}\right]dp &= \int_0^1 \mathbb{1}\{Q_{Y(d)}(p)\le\theta\}\,\mathbb{E}\left[\frac{\mathbb{1}\{D=d\}}{\pi_d(\boldsymbol{X};\zeta)}\right]dp\\ &= \int_0^1 \mathbb{1}\{Q_{Y(d)}(p)\le\theta\}\,dp\\ &= \tau_d(\theta),\end{aligned}$$

and therefore recovers the correct marginal distribution after integration over p.

5.3.3 Tail modelling and extrapolation

The integrated estimating equation introduced in Sections 5.3.1 and 5.3.2 requires a model-assisted representation of the signal function in the extreme tail, where empirical information becomes scarce. When the target quantile lies above a high threshold u, the indicator components of the signal function cannot be reliably evaluated using observed data alone. In this regime, tail extrapolation is unavoidable. The TIEE framework addresses this challenge by embedding an extreme value model directly into the estimating equation, ensuring that extrapolation is both statistically principled and fully integrated with causal identification.

Throughout, we assume that the conditional distribution of the potential outcome $Y(d)$ belongs to a maximum domain of attraction, as formalised in Assumption 5.2. Under this condition, exceedances above a sufficiently high threshold u are asymptotically described by the Generalised Pareto Distribution (GPD). Specifically, for $y > 0$,

$$\Pr(Y(d) - u > y \mid Y(d) > u) = \left(1 + \frac{\xi_d\, y}{\sigma_d}\right)^{-1/\xi_d},$$

where $\sigma_d > 0$ is a scale parameter and $\xi_d \in \mathbb{R}$ is the shape parameter governing tail heaviness. This representation provides the standard peaks-over-threshold approximation for the conditional tail distribution.

The GPD model induces a direct relationship between an extreme quantile level τ and the threshold u. Writing $p_u = \Pr(Y(d) \le u)$, for any quantile satisfying $Q_{Y(d)}(\tau) > u$ we have

$$Q_{Y(d)}(\tau) = u + \frac{\sigma_d}{\xi_d}\left[\left(\frac{1-p_u}{1-\tau}\right)^{\xi_d} - 1\right].$$

This expression forms the basis for extrapolating quantiles beyond the observed support of the data and motivates the effective tail probability $\tau_{\text{eff}} = (\tau - p_u)/(1 - p_u)$ introduced in Section 5.3.1.

EVT enters the TIEE estimating equation through this tail extrapolation. The threshold probability p_u and the GPD scale and shape parameters (σ_d, ξ_d) are estimated from the exceedances and included in the nuisance vector ζ. Replacing them by $(\hat{p}_u, \hat{\sigma}_d, \hat{\xi}_d)$ changes the reconstructed tail outcomes in (5.15) and propagates GPD estimation error into the empirical moment $\mathbf{\Phi}_{d,n}$. The consistency argument requires these nuisance estimators to converge under Assumption 5.4(3), and the asymptotic argument accounts for their first-order contribution through the functional delta method. Thus, the framework incorporates uncertainty from estimating the GPD parameters.

Within the TIEE framework, the GPD does not appear as a standalone plug-in estimator. Instead, it enters through the signal function used in the estimating equation. In the body of the distribution, the signal relies on empirical indicators. In the tail region, these indicators are replaced by model-assisted counterparts derived from the fitted GPD. This substitution ensures continuity in the probability scale between the body and tail components of the integrated estimating equation.

Concretely, for $\theta > u$, the indicator $\mathbb{1}\{Y \le \theta\}$ appearing in the signal function is replaced by a model-assisted version based on the conditional tail distribution. For the inverse probability weighted signal introduced in Section 5.3.2, this yields

$$s_d(W, \theta, \zeta) = \frac{\mathbb{1}\{D = d\}}{\pi_d(X; \zeta)}\,\mathbb{1}\{\tilde{Y}(d) \le \theta\},$$

where $\tilde{Y}(d)$ denotes a reconstructed potential outcome generated from the fitted GPD tail model. A convenient representation is

$$\tilde{Y}(d) = u + \frac{\hat{\sigma}_d}{\hat{\xi}_d}\left[\left(\frac{1-\hat{p}_u}{1-U}\right)^{\hat{\xi}_d} - 1\right], \qquad U \sim \text{Uniform}(0,1), \tag{5.15}$$

or, equivalently, by transforming empirical residuals from the fitted GPD. This construction preserves the probabilistic interpretation of the signal while allowing evaluation of the estimating equation at quantile levels far beyond the observed data. In practice, $\tilde{Y}(d)$ can be obtained using regression models that incorporate observed covariates to adjust for confounding and enhance efficiency. In Sections 5.5 and 5.6 (simulation and empirical application, respectively), we implement this step by fitting such models to the available covariates.

By embedding the GPD tail model within the integrated estimating equation, we avoid the sequential estimate-then-invert strategy commonly used in plug-in extreme quantile estimators. Instead, the extreme quantile is identified as the solution to a single global moment condition that combines causal identification and tail extrapolation. This approach improves numerical stability and ensures that the resulting estimator respects the asymptotic structure imposed by extreme value theory.

The tail calibration described here completes the specification of the TIEE estimating equation. The next section discusses numerical solution and implementation of the resulting estimator.

5.3.4 Numerical solution for the TIEE estimator $\hat{\theta}_{d,\tau}$

The TIEE estimator $\hat{\theta}_{d,\tau}$ is defined as the solution to the empirical analogue of the integrated estimating equation introduced in Section 5.3.1. Specifically, letting

$$\Phi_{d,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_{\text{TIEE},d}(W(i),\theta,\tau,\hat{\zeta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^1 s_{\text{TIEE},d}(W(i),\theta,\hat{\zeta},p)\,dp - \tau\right\}, \tag{5.16}$$

The centring by τ follows directly from the defining relation $g_{\text{TIEE},d}(W,\theta,\tau,\zeta) = S_d(W,\theta) - \tau$ used in the TIEE formulation. Hence the zero-root equation is equivalent to setting the uncentred empirical integrated signal equal to τ. The estimator therefore satisfies $\Phi_{d,n}(\hat{\theta}_{d,\tau}) = 0$. In practice, the integral over $p \in [0,1]$ is approximated numer-

ically using a finite grid. Let $0 = p_0 < p_1 < \cdots < p_K = 1, \qquad p_k - p_{k-1} = \Delta p$, with K sufficiently large. The empirical estimating equation is then approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{k=1}^{K} s_{\mathrm{TIEE},d}(W(i),\theta,\hat{\zeta},p_k)\,(p_k - p_{k-1}) - \tau\right\} = 0. \tag{5.17}$$

In all numerical experiments reported in this chapter, we set $K = 800$ for $n = 1000$ and $K = 2000$ for $n = 5000$, which provides a good balance between numerical accuracy and computational cost. A sensitivity analysis with respect to K is presented in Section 5.5.4.2.

A direct solution of (5.17) using Newton-type methods is often unstable. The difficulty arises because the estimating function contains indicator components inherited from the signal function, which render $\Phi_{d,n}(\theta)$ non-smooth in θ. This issue is particularly pronounced in the extreme tail, where only a small number of observations influence the estimating equation. Following the approach of Peng and Fine (2009) and Ma et al. (2023), we therefore recast the estimating equation as an equivalent convex optimisation problem. To this end, we define an objective function $L_{d,n}(\theta)$ whose subgradient with respect to θ is proportional to $\Phi_{d,n}(\theta)$. A convenient choice is

$$L_{d,n}(\theta) = -\sum_{i=1}^{n}\sum_{k=1}^{K}(p_k - p_{k-1})\,\tilde{s}_{\mathrm{TIEE},d}(W_i,\theta,\hat{\zeta},p_k) + |R^* - \theta|\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{\mathrm{eff}}\,(p_k - p_{k-1}),$$

where $\tilde{s}_{\mathrm{TIEE},d}(W,\theta,\zeta,p_k)$ is an antiderivative of the signal function with respect to θ at p_k, i.e., it satisfies

$$\frac{\partial}{\partial\theta}\tilde{s}_{\mathrm{TIEE},d}(W,\theta,\zeta,p_k) = s_{\mathrm{TIEE},d}(W,\theta,\zeta,p_k),$$

and R^* is a sufficiently large constant ensuring boundedness of the optimisation domain. The TIEE estimator is then obtained as

$$\widehat{\theta}_{d,\tau} = \widehat{Q}_{Y(d)}(\tau) = \arg\min_{\theta\in\Theta} L_{d,n}(\theta).$$

This L_1-type convex formulation guarantees numerical stability and uniqueness of the solution, and it can be solved efficiently (using standard convex optimisation routines) and reliably even when the target quantile lies far beyond the range of the observed data since empirical indicators in the tail region are replaced with model-assisted counterparts, as described in 5.3.3.

5.4 Theoretical properties

In this section, we establish the theoretical foundations of the TIEE framework. Our study centres on the TIEE estimating function $g_{\text{TIEE},d}(W, \theta, \tau, \zeta)$ defined in (5.13), and its associated population integrated moment $\Phi_d(\theta) = \mathbb{E}[g_{\text{TIEE},d}(W, \theta, \tau, \zeta)]$. Note that by construction, $\Phi_d(\theta) = S_d(\theta) - \tau$, where $S_d(\theta) = \mathbb{E}[S_d(W, \theta)]$ is the integrated signal defined in (5.11). Thus, $\Phi_d(\theta)$ is simply the centred version of $S_d(\theta)$. We assume that $\theta \in \Theta$, where $\Theta \subset \mathbb{R}$ is a compact parameter space.

5.4.1 Identification and uniqueness

Identification hinges on two crucial properties of the integrated moment function $g_{\text{TIEE},d}$. First, that it correctly recovers the quantile condition when evaluated at $\theta_{d,\tau}$, and second, that it is strictly monotonic in θ. We require the following assumptions.

Assumption 5.3 (Basic Regularity).

1. Y is a real-valued random variable with continuous marginal distribution function F_Y.
2. The true quantile $\theta_{d,\tau}$ is uniquely defined by $F_{Y_d}(\theta_{d,\tau}) = \tau$, where $\tau \in (0, 1)$ is the target quantile level.
3. For every (W, τ, ζ), the function $g_{\text{TIEE},d}(W, \theta, \tau, \zeta)$ is measurable and *strictly increasing* in θ.
4. The mapping $\Phi_d : \theta \mapsto \mathbb{E}[g_{\text{TIEE},d}(W, \theta, \tau, \zeta)]$ is well-defined on Θ and differentiable in θ.

Assumption 5.3(1)–(2) are standard in quantile estimation. Assumption 5.3(3) is the key monotonicity property that ensures $\Phi_d(\theta)$ crosses zero only once; it holds for all signal functions considered in Section 5.3.2 because they are based on indicator functions. Assumption 5.3(4) ensures that derivatives are well-behaved, which is required for asymptotic analysis.

Theorem 5.1 (Identification and Uniqueness)**.** *Under Assumption 5.3,* $\Phi_d(\theta_{d,\tau}) = 0$ *and* $\Phi_d(\theta)$ *is strictly increasing in* θ*. Consequently, the equation* $\Phi_d(\theta) = 0$ *admits the* unique *solution* $\theta = \theta_{d,\tau}$*.*

Proof. By the unbiasedness property of the signal function (see (5.12)), we have

$$\Phi_d(\theta_{d,\tau}) = \mathbb{E}\big[g_{\mathrm{TIEE},d}(W,\theta,\tau,\zeta)\big] = \mathbb{E}\left[\int_0^1 \big\{s(W,\theta_{d,\tau},\zeta,p) - \tau_{\mathrm{eff}}\big\}\,\mathrm{d}p\right] = 0.$$

Assumption 5.3(3) implies that for every fixed (W,τ,ζ),

$$\theta_{1,\tau} < \theta_2 \quad\Longrightarrow\quad g_{\mathrm{TIEE},d}(W,\theta_{1,\tau},\tau,\zeta) < g_{\mathrm{TIEE},d}(W,\theta_2,\tau,\zeta)$$

Taking expectations preserves the strict inequality, so

$$\Phi_d(\theta_{1,\tau}) < \Phi_d(\theta_2) \quad \text{whenever } \theta_{1,\tau} < \theta_2.$$

Because Φ is strictly increasing and $\Phi_d(\theta_{d,\tau}) = 0$, the equation $\Phi_d(\theta) = 0$ has at most one root, which must coincide with $\theta_{d,\tau}$. □

5.4.2 Consistency of the TIEE estimator

Having established that the true extreme quantile $\theta_{d,\tau}$ is uniquely identified as the root of the population integrated moment $\Phi_d(\theta) = 0$, we now turn to the convergence of its empirical counterpart. The TIEE estimator $\hat{\theta}_{d,\tau}$ is obtained by solving the empirical inverse estimating equation $\Phi_{d,n}(\theta) = 0$, where $\Phi_{d,n}(\theta)$ is defined in (5.16) as the sample average of the integrated moment condition. Note that $\Phi_{d,n}(\theta)$ incorporates estimated nuisance parameters $\hat{\zeta}$ (propensity scores, GPD parameters, etc.), making $\hat{\theta}_{d,\tau}$ a Z-estimator. Consistency relies on the uniform convergence of the empirical functional $\Phi_n(\theta)$ to its population counterpart $\Phi_d(\theta)$. For this, we require the following assumptions.

Assumption 5.4 (Regularity for Consistency).

1. The covariate $\boldsymbol{X}$ has compact support.
2. The mapping $\theta \mapsto \mathbb{1}\{Q_{Y(d)}(p) \leq \theta\}$ is monotonic and piecewise constant for fixed p.
3. The nuisance parameter estimates $\hat{\zeta}$ converge in probability to the true parameters ζ, that is, $\hat{\zeta} \xrightarrow{P} \zeta$.
4. For the inverse-probability-weighted signal, overlap holds uniformly: there exists $\varepsilon > 0$ such that $\pi_d(\boldsymbol{X}; \zeta) \geq \varepsilon$ almost surely for $d \in \{0, 1\}$.

Assumption 5.4(1) is standard in causal inference and provides the compactness required for uniform convergence results. Assumption 5.4(2) reflects the step-function nature of quantile functions and is automatically satisfied for continuous distributions. Assumption 5.4(3) requires consistent estimation of nuisance parameters; for the propensity score, this follows from standard semiparametric theory (Hirano et al., 2003), whilst for GPD parameters, this follows from maximum likelihood theory (De Haan and Ferreira, 2007).

Lemma 5.1 (Glivenko–Cantelli property of the integrated signal)**.** *For $\theta \in \Theta$, let $S_d(W, \theta)$ be defined as in (5.10), with $s_{\mathrm{TIEE},d}$ equal to either the naïve signal in (5.14) or the inverse-probability-weighted signal in Section 5.3.2. Under Assumption 5.4, the class $\mathcal{F} = \{S_d(\cdot, \theta) : \theta \in \Theta\}$ is Glivenko–Cantelli.*

Proof. For each fixed $p \in (0, 1)$, the threshold indicators $\{\mathbb{1}\{Q_{Y(d)}(p) \leq \theta\} : \theta \in \Theta\}$ form a VC-subgraph class. The naïve signal is therefore Glivenko–Cantelli. For the inverse-probability-weighted signal, the indicator is multiplied by the fixed measurable weight $\mathbb{1}\{D = d\}/\pi_d(\boldsymbol{X}; \zeta)$. Assumption 5.4(4) bounds this weight by $1/\varepsilon$, so multiplication preserves the Glivenko–Cantelli property. In either case the signal has an integrable envelope uniformly in p. Integration over $p \in (0, 1)$ is a bounded linear operation and hence preserves the Glivenko–Cantelli property (Van Der Vaart and Wellner, 2023, Sections 2.4 and 2.6). Therefore,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} S_d(W_i, \theta) - \mathbb{E}\big[S_d(W, \theta)\big] \right| \xrightarrow{P} 0.$$

Thus, $\mathcal{F}$ is Glivenko–Cantelli. □

Theorem 5.2 (Consistency)**.** *Under Assumptions 5.3 and 5.4, and provided that the associated integrated-signal class is Glivenko–Cantelli, the TIEE estimator $\hat{\theta}_{d,\tau}$ that solves the empirical estimating equation $\Phi_{d,n}(\theta) = 0$ is consistent for the true extreme quantile $\theta_{d,\tau}$, i.e., $\hat{\theta}_{d,\tau} \xrightarrow{P} \theta_{d,\tau}$.*

Proof. The consistency of $\hat{\theta}_{d,\tau}$, defined as the root of the empirical estimating equation $\Phi_n(\theta) = 0$, follows from the general consistency theory for Z–estimators (Van der Vaart, 2000, Theorem 5.9). This requires verifying unique identification and uniform convergence. Unique identification holds by Theorem 5.1. For uniform convergence, note that by the stated Glivenko–Cantelli condition the class $\mathcal{F} = \{S_d(\cdot, \theta) : \theta \in \Theta\}$ is Glivenko–Cantelli. Since $g_{\mathrm{TIEE},d}(W, \theta, \tau) = S_d(W, \theta) - \tau$ (see (5.13)), the class $\{g_{\mathrm{TIEE},d}(\cdot; \theta, \tau) : \theta \in \Theta\}$ is also Glivenko–Cantelli. Consequently,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} g_{\mathrm{TIEE},d}(W_i, \theta, \tau) - \mathbb{E}[g_{\mathrm{TIEE},d}(W, \theta, \tau)] \right| \xrightarrow{P} 0.$$

By Assumption 5.4(3), replacing ζ by $\hat{\zeta}$ introduces an $o_P(1)$ error uniformly in θ, yielding

$$\sup_{\theta \in \Theta} \left| \Phi_{d,n}(\theta) - \Phi_d(\theta) \right| \xrightarrow{P} 0.$$

Finally, since $\Phi_d(\theta)$ admits a unique zero at $\theta_{d,\tau}$ and is strictly monotone in θ (Theorem 5.1), uniform convergence of $\Phi_{d,n}$ to Φ_d implies that any solution $\hat{\theta}_{d,\tau}$ to $\Phi_{d,n}(\theta) = 0$ satisfies $\hat{\theta}_{d,\tau} \xrightarrow{P} \theta_{d,\tau}$ (Van der Vaart, 2000, Theorem 5.9). □

Theorem 5.2 establishes that the TIEE estimator is consistent even when the target quantile τ is extreme and lies beyond the range of observed data. This result is non-trivial because it holds despite the vanishing density at $\theta_{d,\tau}$, a consequence of the EVT-based regularisation embedded in the integrated moment condition.

5.4.3 Asymptotic normality and efficiency

Building on the consistency result in Theorem 5.2, we derive the asymptotic distribution of the TIEE estimator via a first-order Taylor expansion of the empirical estimating equation $\Phi_n(\hat{\theta}_{d,\tau}) = 0$ around the target quantile $\theta_{d,\tau}$. This expansion requires sufficient smoothness of the TIEE moment functional and finiteness of the second moment of the associated score function.

Assumption 5.5 (Regularity for Asymptotic Normality).

1. The function $\Phi_d(\theta)$ is continuously differentiable in a neighbourhood of $\theta_{d,\tau}$ with $\Phi'_d(\theta_{d,\tau}) \neq 0$.
2. The nuisance parameter estimates $\hat{\zeta}$ satisfy $\sqrt{n}(\hat{\zeta} - \zeta) = O_p(1)$, and $\Phi_{d,n}(\theta)$ is pathwise differentiable with respect to ζ.
3. The variance $\sigma^2_{d,\tau} := \mathrm{Var}[g_{\mathrm{TIEE},d}(W, \tau, \theta_{d,\tau}, \zeta)]$ is finite.

Assumption 5.5(1) ensures that the derivative of $\Phi_d(\theta)$ exists and is non-zero at $\theta_{d,\tau}$, which is required for the implicit function theorem to apply. Under our signal functions, $\Phi'(\theta_{d,\tau}) = f_{Y(d)}(\theta_{d,\tau})$, the density of the potential outcome at the extreme quantile. Whilst this density may be small, it remains strictly positive under Assumption 5.5(1). Assumption 5.5(2) controls the rate at which nuisance parameters are estimated, which is standard in semiparametric theory and holds for logistic regression propensity scores and maximum likelihood GPD estimates. Assumption 5.5(3) is a moment condition ensuring the central limit theorem applies.

Theorem 5.3 (Asymptotic Normality)**.** *Under Assumptions 5.3, 5.4, and 5.5, the TIEE estimator $\hat{\theta}_{d,\tau}$ is asymptotically normal:*

$$\sqrt{n}(\hat{\theta}_{d,\tau} - \theta_{d,\tau}) \xrightarrow{d} \mathcal{N}(0, V_{d,\tau}), \quad \textit{with} \quad V_{d,\tau} = \frac{\sigma^2_{d,\tau}}{[\Phi'_d(\theta_{d,\tau})]^2}. \tag{5.18}$$

A complete proof, which includes accounting for the estimation of nuisance parameters $\hat{\zeta}$ via the functional delta method (Van der Vaart, 2000), is presented in the Appendix.

The asymptotic variance in (5.18) highlights several distinctive features of the TIEE estimator. First, unlike classical empirical quantiles, the variance remains finite even in extreme regimes where $f_Y(\theta_{d,\tau}) \to 0$ as $\tau \to 1$, which is a direct consequence of the EVT-based regularisation. Moreover, integrating the estimating equation over quantile levels

in (5.11) allows the estimator to borrow strength from neighbouring quantiles, stabilising inference in the tail. This construction also reconciles our framework with standard extreme value limits. While empirical quantiles converge to non-Gaussian limits when the effective tail sample size is bounded ($n(1-\tau_n) \to d > 0$), the TIEE estimator attains asymptotic normality by operating under an intermediate sequence regime. The GPD tail parameters (σ, ξ) are estimated from k exceedances with $k \to \infty$ and $k/n \to 0$, yielding $\sqrt{k}$-consistent estimates. Since the extreme quantile estimator is obtained as a smooth function of these parameters through the extrapolation formula in (5.15), asymptotic normality follows via the Delta method.

The variance $V_{d,\tau}$ further depends on the derivative of the population moment function, $\Phi'_d(\theta_{d,\tau})$. Under regularity conditions permitting differentiation under the expectation,

$$\Phi'(\theta_{d,\tau}) = \mathbb{E}\left[\frac{\partial\, g_{\mathrm{TIEE},d}(W, \tau, \theta_{d,\tau}, \zeta)}{\partial\theta}\right].$$

For both signal functions considered in Section 5.3.2, this derivative coincides with the density of the potential outcome evaluated at the target quantile, $\Phi'_d(\theta) = f_{Y(d)}(\theta)$, by an application of Leibniz's rule.

Finally, the variance $\sigma^2_{d,\tau}$ accounts for uncertainty arising from the estimation of the nuisance parameters ζ. Under Assumption 5.5(2), this contribution can be characterised through the efficient influence function (Tsiatis, 2006). In particular, when the propensity score is correctly specified, the IPW-based TIEE estimator is asymptotically efficient for extreme quantile treatment effect estimation.

5.4.4 Inference for the extreme quantile treatment effect $\delta(\tau)$

Based on the results from Sections 5.4.1-5.4.3, we here derive an expression for the asymptotic distribution and derive confidence intervals for the extreme quantile treatment effect $\delta(\tau) = \theta_{1,\tau} - \theta_{0,\tau}$ (as defined in (5.1)). We start by deriving the joint asymptotic distribution of $(\theta_{1,\tau}, \theta_{0,\tau})^\top$. A first order Taylor expansion of the empirical estimating equation $\Phi_{d,n}(\widehat{\theta}_{d,\tau}) = 0$ around $\theta_{d,\tau}$ gives

$$0 = \Phi_{d,n}(\widehat{\theta}_{d,\tau}) \approx \Phi_{d,n}(\theta_{d,\tau}) + \Phi'_{d,n}(\theta_{d,\tau})(\hat{\theta}_{d,\tau} - \theta_{d,\tau}), \quad d = 0, 1,$$

which means that

$$\sqrt{n}(\hat{\theta}_d - \theta_{d,\tau}) \approx -\frac{\sqrt{n}\Phi_{d,n}(\theta_{d,\tau})}{\Phi'_{d,n}(\theta_{d,\tau})}.$$

By consistency and uniform convergence, $\Phi'_{d,n}(\theta_{d,\tau}) \xrightarrow{P} \Phi'_d(\theta_{d,\tau})$, so asymptotically,

$$\sqrt{n}(\hat{\theta}_d - \theta_{d,\tau}) = -\frac{\sqrt{n}\Phi_{d,n}(\theta_{d,\tau})}{\Phi'_d(\theta_{d,\tau})} + o_p(1).$$

Now, stacking the two treatment groups together, we get

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_{1,\tau} - \theta_{1,\tau} \\ \hat{\theta}_{0,\tau} - \theta_{0,\tau}\end{pmatrix} = -D^{-1}\sqrt{n}\begin{pmatrix}\Phi_{n,1}(\theta_{1,\tau}) \\ \Phi_{n,0}(\theta_{0,\tau})\end{pmatrix} + o_p(1), \qquad D = \operatorname{diag}\big(\Phi'_1(\theta_{1,\tau}), \Phi'_0(\theta_{0,\tau})\big).$$

By the multivariate central limit theorem,

$$\sqrt{n}\begin{pmatrix}\Phi_{n,1}(\theta_{1,\tau}) \\ \Phi_{n,0}(\theta_{0,\tau})\end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma_\Phi),$$

where

$$\Sigma_\Phi = \begin{pmatrix}\Sigma_{11} & \Sigma_{10} \\ \Sigma_{10} & \Sigma_{00}\end{pmatrix} = \begin{pmatrix}\operatorname{Var}(g_1(W, \theta_{1,\tau})) & \operatorname{Cov}(g_1(W, \theta_{1,\tau}), g_0(W, \theta_{0,\tau})) \\ \operatorname{Cov}(g_1(W, \theta_{1,\tau}), g_0(W, \theta_{0,\tau})) & \operatorname{Var}(g_0(W, \theta_{0,\tau}))\end{pmatrix}. \tag{5.19}$$

In (5.19), we have reduced notation clutter by denoting $g_d(W, \theta_{d,\tau})) = g_{\text{TIEE},d}(W, \theta, \tau, \zeta)$, so $\operatorname{Var}(g_d(W, \theta_{d,\tau}))$ is the variance of the integrated estimating function for treatment d, which can be estimated using the sample variance:

$$\hat{\Sigma}_{dd} = \frac{1}{n}\sum_{i=1}^{n} g_d(W(i), \hat{\theta}_{1,\tau})^2 - \left(\frac{1}{n}\sum_{i=1}^{n} g_d(W(i), \hat{\theta}_{1,\tau})\right)^2 \approx \frac{1}{n}\sum_{i=1}^{n} g_d(W(i), \hat{\theta}_{1,\tau})^2, \quad d = 0, 1. \tag{5.20}$$

The approximation in (5.20) comes from the fact that $\widehat{\theta}_{d,\tau}$ solves $\Phi_{d,n}(\widehat{\theta}_{d,\tau}) = 0$, therefore the sample mean is close to 0. The covariance term Σ_{10} captures the dependence between the estimating equations for the two treatment groups. When treatment and control samples are independent (e.g., in randomised experiments with non-overlapping units), $\Sigma_{10} = 0$. However, in observational studies with IPW adjustment, $\Sigma_{10} \neq 0$ in general because both estimators use the same propensity score model and the same covariate

distribution. In that case, the covariance can be estimated by

$$\hat{\Sigma}_{10} = \frac{1}{n}\sum_{i=1}^{n} g_1(W(i), \hat{\theta}_{1,\tau}) g_0(W(i), \hat{\theta}_{0,\tau}) - \left(\frac{1}{n}\sum_{i=1}^{n} g_1(W(i), \hat{\theta}_{1,\tau})\right)\left(\frac{1}{n}\sum_{i=1}^{n} g_0(W(i), \hat{\theta}_{0,\tau})\right).$$

By Slutsky's theorem, the joint re-scaled asymptotic distribution of $(\widehat{\theta}_{1,\tau}, \widehat{\theta}_{0,\tau})^\top$ is given by

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_{1,\tau} - \theta_{1,\tau} \\ \hat{\theta}_{0,\tau} - \theta_{0,\tau}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, D^{-1}\Sigma_\Phi D^{-1}\right).$$

The asymptotic distribution of $\hat{\delta}(\tau)$ is therefore $\sqrt{n}\left(\hat{\delta}(\tau) - \delta(\tau)\right) \xrightarrow{d} \mathcal{N}\left(0,\ \sigma_\delta^2\right)$, where the asymptotic variance $\sigma_\delta^2 = (1, -1)\, D^{-1}\Sigma_\Phi D^{-1}\, (1, -1)^\top$ simplifies to

$$\sigma_\delta^2 = \frac{\Sigma_{11}}{[\Phi_1'(\theta_{1,\tau})]^2} + \frac{\Sigma_{00}}{[\Phi_0'(\theta_{0,\tau})]^2} - \frac{2\Sigma_{10}}{\Phi_1'(\theta_{1,\tau})\Phi_0'(\theta_{0,\tau})}. \tag{5.21}$$

A consistent estimator $\hat{\sigma}_\delta^2$ of (5.21) is obtained by substituting the terms on the right-hand side by sample analogues. The terms in the denominators can either be estimated via numerical differentiation or kernel density estimation of $f_{Y(d)}(\hat{\theta}_{d,\tau})$ (since $\Phi_d'(\theta_{d,\tau}) = f_{Y(d)}(\hat{\theta}_{d,\tau})$). An asymptotically valid $(1-\alpha)$ confidence interval for $\widehat{\delta}(\tau)$ is thefore given by

$$\widehat{\delta}(\tau) \pm z_{1-\alpha/2}\frac{\hat{\sigma}_\delta}{\sqrt{n}},$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution.

5.5 Simulation study

To evaluate the finite-sample performance of the proposed TIEE framework, we conduct a comprehensive simulation study with three objectives: (1) to compare our TIEE framework using inverse probability weighting (TIEE-IPW) against state-of-the-art competing methods, including the Zhang-Firpo estimator proposed in Zhang (2018a), the Hill causal estimator in Deuber et al. (2024), and the Pickands estimator in Zhang (2018b); (2) to assess estimator behaviour across different tail heaviness regimes (heavy-tailed versus light-tailed distributions); and (3) to empirically verify the robustness of the TIEE-IPW estimator under propensity score misspecification.

We consider two sample sizes, $n \in \{1000, 5000\}$, and report the bias and mean squared error (MSE) based on 1000 independent replications. The target levels are specified through $(1 - \tau_n) \in \{5/n, 1/n, 5/(n \log n)\}$. These choices correspond to extrapolative regimes in which the expected number of observations beyond the target quantile, $n(1-\tau_n)$, is bounded (equal to 5 or 1) or vanishes (equal to $5/\log n$). The intermediate regime $n(1 - \tau_n) \to \infty$ is not included because the experiment focuses on quantiles at or beyond the empirically supported tail.

5.5.1 Data-generating process

The data-generating process is consistent across all scenarios. We generate n independent and identically distributed observations $W(i) = (Y(i), D(i), \boldsymbol{X}(i))$. The covariates $\boldsymbol{X}(i)$ consist of a constant intercept and a continuous variable $X(i) \sim \text{Uniform}(-1, 1)$, such that $\boldsymbol{X}(i) = (1, X(i))^\top$. The propensity score for the binary treatment $D = 1$ is defined by a quadratic function of the covariate x:

$$\pi(\boldsymbol{x}) = \Pr(D = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.5x^2 + 0.25. \tag{5.22}$$

The treatment assignment $D(i)$ is then generated as $D(i) \mid \boldsymbol{X}(i) \sim \text{Ber}(\pi(\boldsymbol{X}(i)))$. The potential outcomes $Y(1)$ and $Y(0)$ are specified according to the tail regime scenarios detailed below.

Following Deuber et al. (2024), we set the intermediate quantile level associated with the GPD threshold as $(1-p_u) = k/n$ with $k = n^{0.65}$. Whilst this choice may not be optimal in all settings, it provides a reasonable benchmark for comparison. A comprehensive sensitivity analysis regarding the choice of k is presented in Section 5.5.4. The reconstructed potential outcome tail distribution is obtained via GPD regression on exceedances above a high threshold, where the GPD parameters are modelled as functions of covariates.

5.5.2 Heavy-tailed scenarios

This set of scenarios focuses on cases where the underlying potential outcome distributions F_{Y_d} belong to the Fréchet maximum domain of attraction, characterised by heavy tails. The three scenarios vary the heterogeneity of the EVIs between the treatment and control groups. Following Deuber et al. (2024), the potential outcomes are generated from the following models:

$$M_1^{(\mathrm{H})}: \begin{cases} Y(1) = 5S(1+X), \\ Y(0) = S(1+X), \end{cases}$$

$$M_2^{(\mathrm{H})}: \begin{cases} Y(1) = C_2 \exp(X), \\ Y(0) = C_3 \exp(X), \end{cases}$$

$$M_3^{(\mathrm{H})}: \begin{cases} Y(1) = P_{1.75+X,2}, \\ Y(0) = P_{1.75+X,1}, \end{cases}$$

where X is defined as above, S follows a Student-t distribution with 3 degrees of freedom, C_s is Fréchet distributed with shape parameter s, location 0, and scale 1, and $P_{a,b}$ is Pareto distributed with shape parameter a and scale b. The associated EVIs for each scenario are $\gamma_1 = \gamma_0 = 1/3$ for model $M_1^{(\mathrm{H})}$, $\gamma_1 = 1/2$ and $\gamma_0 = 1/3$ for model $M_2^{(\mathrm{H})}$, and $\gamma_1 = \gamma_0 = 4/7$ for model $M_3^{(\mathrm{H})}$.

5.5.2.1 Results

Table 5.1 presents the estimation performance under heavy-tailed scenarios, where the TIEE-IPW estimator demonstrates superior performance across nearly all scenarios. Specifically, it shows consistently lower MSE compared to the Zhang-Firpo and Pickands estimators across the target probability sequences, although Zhang-Firpo often attains a smaller absolute bias, particularly for $n = 1000$ and in the light-tailed scenarios. Whilst the Causal Hill estimator performs reasonably well in the least extrapolative setting

($(1-\tau_n) = 5/n$), the TIEE-IPW estimator maintains robust performance across all three target sequences, including the very extreme tail ($(1-\tau_n) = 5/(n\log n)$). The performance advantage of TIEE-IPW becomes more pronounced in the most extreme tail regions, where extrapolation-based methods face greater challenges.

The Pickands estimator exhibits substantially larger bias and variance across all scenarios, particularly in the very extreme tail regime. This is consistent with the known higher variability of the Pickands estimator in finite samples. The Zhang-Firpo estimator shows competitive performance in the least extrapolative setting but deteriorates notably in Scenario $M_2^{(\mathrm{H})}$, where the heterogeneity in EVIs between treatment and control groups is most pronounced.

Figure 5.1 reports the empirical coverage of the 90% intervals across the three heavy-tailed designs for both $n = 1000$ and $n = 5000$. Both the Causal Hill estimator and the proposed TIEE method exhibit stable coverage behaviour across settings, reflecting their compatibility with Pareto-type tails. By contrast, the Zhang estimator shows systematic under-coverage, with the deviation from the nominal level increasing as the target tail probability becomes more extreme. Overall, TIEE achieves the most accurate finite-sample coverage while remaining robust to varying degrees of tail heaviness. Results based on the Pickands estimator are not reported because its finite-sample variability yields extremely unstable interval estimates.

5.5.3 Light-tailed scenarios

Here we focus on scenarios where the underlying distribution F_{Y_d} belong to the Gumbel or Weibull maximum domain of attraction, corresponding to light-tailed distributions. Here we expect EVT-based estimators that primarily target heavy-tailed data (such as the Hill estimator method) to not perform very well. The potential outcomes are generated from

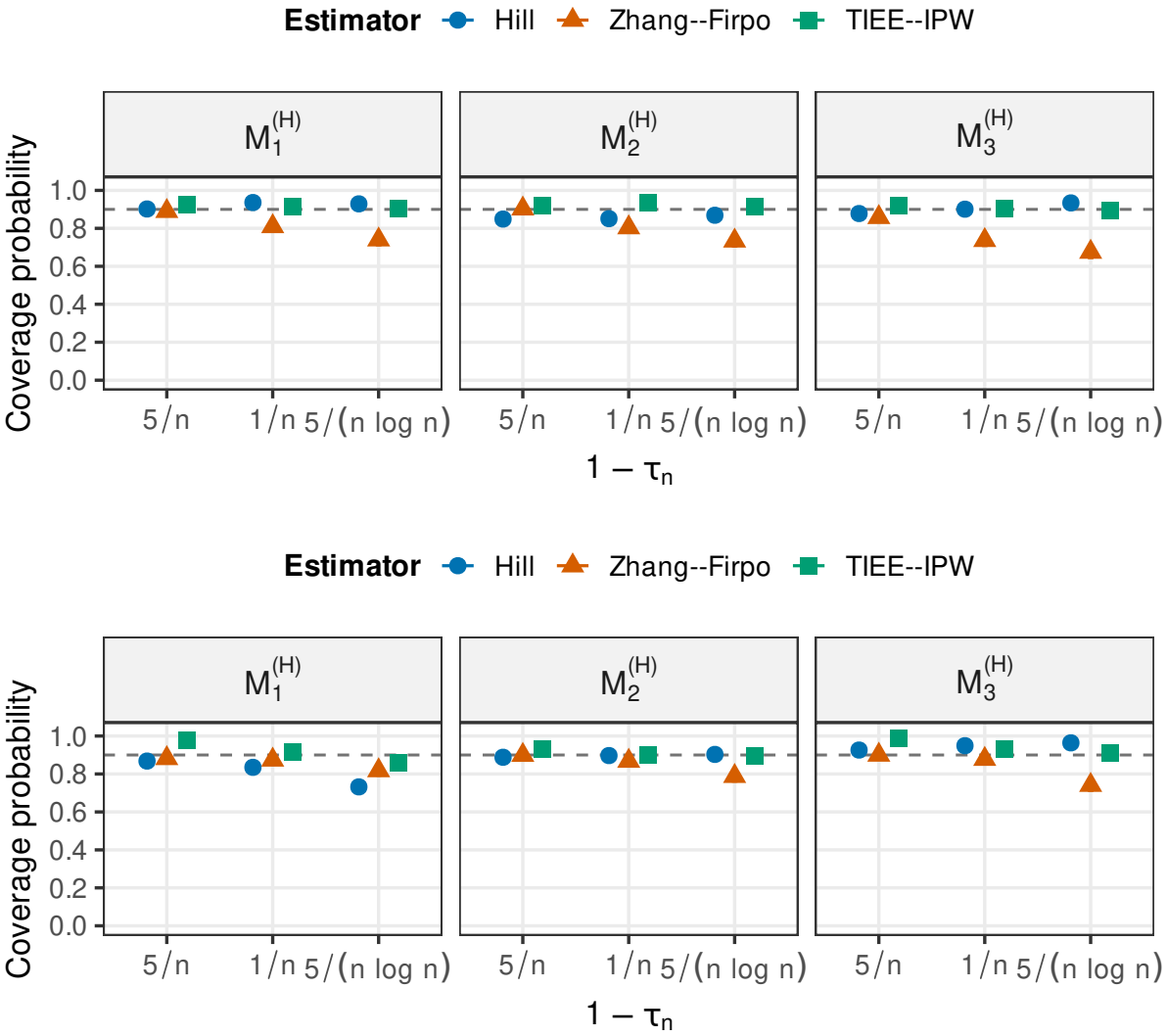

Figure 5.1: Coverage rates of 90% confidence intervals for the extreme quantile treatment effect under the three heavy-tailed scenarios ($M_1^{(\mathrm{H})}$, $M_2^{(\mathrm{H})}$, $M_3^{(\mathrm{H})}$). The upper and lower panels show $n = 1000$ and $n = 5000$, respectively. The comparison includes the Zhang–Firpo, Causal Hill, and TIEE–IPW estimators at target tail probabilities $5/n$, $1/n$, and $5/(n \log n)$. The dashed horizontal line denotes the nominal coverage level. The estimand is $\delta(\tau_n) = Q_{Y(1)}(\tau_n) - Q_{Y(0)}(\tau_n)$ from (5.1); for interval $[L_b, U_b]$ in replicate b, empirical coverage is $B^{-1} \sum_{b=1}^{B} \mathbb{1}\{L_b \le \delta(\tau_n) \le U_b\}$ with $B = 1000$.

Table 5.1: Estimation performance of the Causal Hill, Pickands, Zhang–Firpo, and TIEE–IPW estimators of the EQTE $\delta(\tau_n)$ in (5.1) under the three heavy-tailed scenarios, for $n \in \{1000, 5000\}$ and $(1-\tau_n) \in \{5/(n\log n), 1/n, 5/n\}$. Writing $\delta_0(\tau_n)$ for the data-generating value and $\hat{\delta}_b(\tau_n)$ for replicate b, Bias is $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n) - \delta_0(\tau_n)\}$ and MSE is $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n) - \delta_0(\tau_n)\}^2$, with $B = 1000$. Bold entries identify the smallest MSE within each scenario, sample size, and target level.

			Scenario $M_1^{(\mathrm{H})}$		Scenario $M_2^{(\mathrm{H})}$		Scenario $M_3^{(\mathrm{H})}$	
n	$1-\tau_n$	Method	Bias	MSE	Bias	MSE	Bias	MSE
1000	$5/(n\log n)$	Causal Hill	45.982	4828.099	6.370	1355.930	26.759	3652.382
		Pickands	-52.322	5694.769	42.410	56223.057	63.741	86214.119
		Zhang-Firpo	-3.243	2840.382	0.847	5395.063	-9.793	6191.232
		TIEE-IPW	-6.792	**1513.337**	-16.460	**581.743**	-5.879	**544.757**
	$1/n$	Causal Hill	28.305	2329.556	4.388	844.775	15.660	1872.560
		Pickands	-36.222	6194.962	59.070	110234.161	58.364	101126.968
		Zhang-Firpo	4.117	2846.812	8.653	5469.215	-2.782	6103.073
		TIEE-IPW	-4.997	**1025.129**	-10.776	**390.789**	-4.072	**390.293**
	$5/n$	Causal Hill	4.496	135.273	0.539	55.963	2.142	79.982
		Pickands	-19.168	899.346	6.520	1714.555	4.988	1268.937
		Zhang-Firpo	1.194	174.980	1.183	131.197	3.330	1128.588
		TIEE-IPW	-1.193	**101.772**	-2.066	**52.606**	-1.680	**68.289**
5000	$5/(n\log n)$	Causal Hill	40.165	3652.733	22.858	1537.911	35.070	3914.050
		Pickands	70.235	29395.493	130.364	202295.849	153.724	354024.912
		Zhang-Firpo	33.818	4655.698	42.605	14786.807	39.325	9158.521
		TIEE-IPW	22.842	**876.524**	22.348	**996.323**	21.852	**878.432**
	$1/n$	Causal Hill	31.616	2279.542	18.035	942.085	26.499	2181.550
		Pickands	59.719	15831.284	95.367	86841.475	107.704	139943.738
		Zhang-Firpo	31.558	4507.900	40.327	14597.895	35.425	8866.979
		TIEE-IPW	19.310	**695.274**	18.341	**761.169**	17.668	**632.377**
	$5/n$	Causal Hill	8.553	182.638	5.164	**72.501**	6.128	107.085
		Pickands	26.375	1308.698	20.979	1807.624	19.685	1832.421
		Zhang-Firpo	9.558	272.607	7.824	208.126	8.530	447.514
		TIEE-IPW	7.061	**131.785**	5.947	114.472	5.163	**75.775**

the following three models:

$$M_1^{(\mathrm{L})}: \begin{cases} Y(1) = 5S(1+X), \\ Y(0) = S(1+X), \end{cases}$$

$$M_2^{(\mathrm{L})}: \begin{cases} Y(1) = C_1 \exp(X), \\ Y(0) = C_2 \exp(X), \end{cases}$$

$$M_3^{(\mathrm{L})}: \begin{cases} Y(1) = W_{2+X,2}, \\ Y(0) = W_{3+2X,1}, \end{cases}$$

where X is defined as above, S follows a standard normal distribution, $C_1 \sim \text{Exp}(1)$, $C_2 \sim \text{Exp}(2)$, and $W_{a,b}$ denotes a Weibull distribution with shape parameter a and scale b. The associated EVIs for each scenario are $\gamma_1 = \gamma_0 = 0$ for models $M_1^{(\text{L})}$ and $M_2^{(\text{L})}$ (corresponding to the Gumbel domain of attraction), and $\gamma_1 = -1/(2+X)$ and $\gamma_0 = -1/(3+2X)$ for model $M_3^{(\text{L})}$.

5.5.3.1 Results

Table 5.2 presents the estimation performance under light-tailed scenarios. The TIEE-IPW estimator maintains its superiority across all light-tailed scenarios, achieving the lowest MSE in every case. Overall, we can see that all estimators perform substantially better compared to heavy-tailed scenarios, as evidenced by the markedly lower MSE values. This is expected, as the tails of light-tailed distributions are inherently easier to estimate. For this reason, the relative advantage of TIEE-IPW over competing methods is maintained but less pronounced than in the heavy-tailed case. The Pickands estimator continues to exhibit the poorest performance, with substantially inflated MSE values, particularly in the most extrapolative tail regime.

Figure 5.2 reports the empirical coverage of the 90% intervals across the three light-tailed designs for both $n = 1000$ and $n = 5000$. Overall, TIEE achieves stable coverage close to the nominal level across designs and quantile regimes. Zhang's estimator systematically under-covers, with the discrepancy increasing at more extreme quantiles, consistent with its attenuation under light-tailed behaviour. Causal Hill's estimator shows design-dependent performance. While it substantially under-covers in $M_1^{(\text{L})}$ and $M_3^{(\text{L})}$, particularly at the most extreme quantiles, it attains near-nominal coverage in $M_2^{(\text{L})}$, indicating that classical tail-index methods can perform well in especially benign light-tail settings. In contrast, TIEE provides the most uniform uncertainty quantification across designs, avoiding reliance on a small number of extreme observations and remaining robust to mild-tail behaviour and small exceedance probabilities.

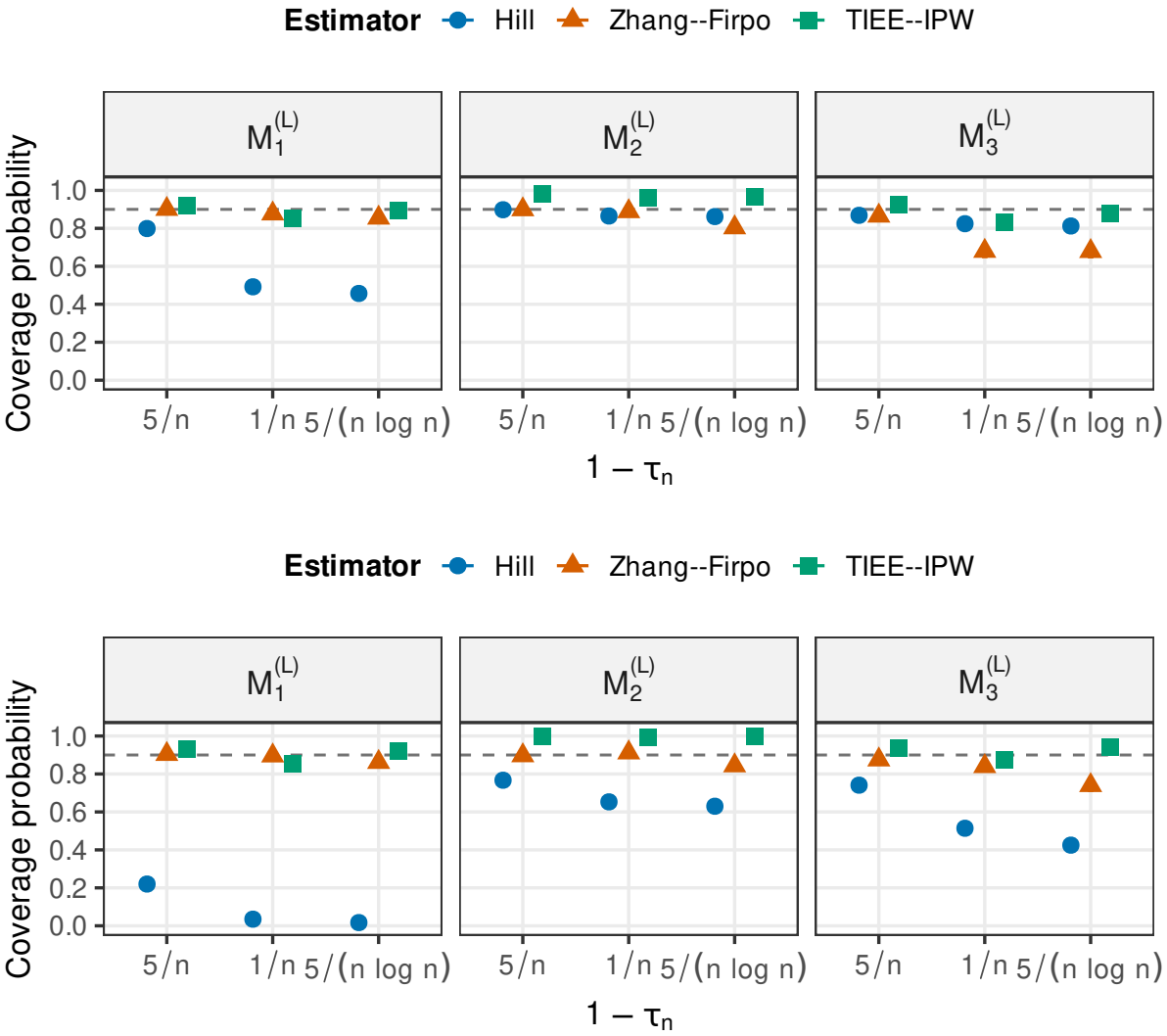

Figure 5.2: Coverage rates of 90% confidence intervals for the extreme quantile treatment effect under the three light-tailed scenarios ($M_1^{(L)}$, $M_2^{(L)}$, $M_3^{(L)}$). The upper and lower panels show $n = 1000$ and $n = 5000$, respectively. The comparison includes the Zhang–Firpo, Causal Hill, and TIEE–IPW estimators at target tail probabilities $5/n$, $1/n$, and $5/(n \log n)$. The dashed horizontal line denotes the nominal coverage level. The estimand is $\delta(\tau_n) = Q_{Y(1)}(\tau_n) - Q_{Y(0)}(\tau_n)$ from (5.1); for interval $[L_b, U_b]$ in replicate b, empirical coverage is $B^{-1} \sum_{b=1}^{B} \mathbb{1}\{L_b \le \delta(\tau_n) \le U_b\}$ with $B = 1000$.

Table 5.2: Estimation performance of the Causal Hill, Pickands, Zhang–Firpo, and TIEE–IPW estimators of the EQTE $\delta(\tau_n)$ in (5.1) under the three light-tailed scenarios, for $n \in \{1000, 5000\}$ and $(1-\tau_n) \in \{5/(n\log n), 1/n, 5/n\}$. Writing $\delta_0(\tau_n)$ for the data-generating value and $\hat{\delta}_b(\tau_n)$ for replicate b, Bias is $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n) - \delta_0(\tau_n)\}$ and MSE is $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n) - \delta_0(\tau_n)\}^2$, with $B = 1000$. Bold entries identify the smallest MSE within each scenario, sample size, and target level.

			Scenario $M_1^{(\mathrm{L})}$		Scenario $M_2^{(\mathrm{L})}$		Scenario $M_3^{(\mathrm{L})}$	
n	$1-\tau_n$	Method	Bias	MSE	Bias	MSE	Bias	MSE
1000	$5/(n\log n)$	Causal Hill	13.488	244.970	4.372	49.431	0.535	0.601
		Pickands	-15.687	482.247	5.818	3791.332	-1.827	15.487
		Zhang-Firpo	-0.337	14.195	0.475	12.403	-0.347	0.307
		TIEE-IPW	-1.286	**10.593**	-0.153	**9.121**	-0.097	**0.252**
	$1/n$	Causal Hill	10.916	163.065	3.459	32.668	0.441	0.447
		Pickands	-15.195	408.884	3.743	2018.179	-1.788	12.386
		Zhang-Firpo	0.406	14.245	0.847	12.895	-0.266	0.257
		TIEE-IPW	-1.144	**8.332**	-0.109	**6.573**	-0.089	**0.196**
	$5/n$	Causal Hill	2.563	12.662	0.734	3.189	0.105	0.083
		Pickands	-11.699	183.880	-0.564	106.240	-1.438	4.721
		Zhang-Firpo	-0.090	4.283	0.058	2.774	-0.015	0.089
		TIEE-IPW	-0.488	**2.484**	0.173	**1.334**	-0.063	**0.050**
5000	$5/(n\log n)$	Causal Hill	17.536	339.969	6.735	71.926	0.797	0.826
		Pickands	-20.353	492.972	-1.767	492.544	-2.236	10.139
		Zhang-Firpo	-0.770	11.042	0.196	13.396	-0.386	0.322
		TIEE-IPW	-1.688	**7.863**	-0.683	**5.600**	-0.173	**0.138**
	$1/n$	Causal Hill	13.458	201.658	4.997	40.885	0.628	0.535
		Pickands	-19.186	430.554	-2.139	293.848	-2.118	8.627
		Zhang-Firpo	0.316	10.549	0.847	14.074	-0.262	0.242
		TIEE-IPW	-1.456	**5.861**	-0.486	**3.930**	-0.150	**0.104**
	$5/n$	Causal Hill	4.903	28.551	1.604	5.565	0.250	0.113
		Pickands	-15.105	259.551	-2.340	64.948	-1.685	4.994
		Zhang-Firpo	-0.169	2.942	-0.116	2.348	-0.044	0.070
		TIEE-IPW	-0.876	**2.081**	-0.076	**1.095**	-0.105	**0.039**

5.5.4 Sensitivity and robustness analysis

We conduct a thorough sensitivity analysis to examine the stability of the TIEE estimator under critical implementation choices, focusing primarily on the extreme value threshold u and robustness to propensity score misspecification.

5.5.4.1 Sensitivity to the threshold u

The TIEE estimator's performance depends on the appropriate selection of the intermediate quantile level p_u associated with the threshold u. This sensitivity study uses the heavy-tailed Pareto design $M_3^{(\mathrm{H})}$ defined in Section 5.5.2. We fix the sample size at $n = 1000$ and consider the target quantile levels τ_n with $(1 - \tau_n) \in \{5/n, 1/n, 5/(n \log n)\}$. We vary the common quantile index over $p_u \in [0.70, 0.95]$ in increments of 0.01; the corresponding outcome threshold is estimated separately within each treatment group, so u ranges from the 70th to the 95th percentile of that group's conditional outcome distribution.

Figure 5.3 reports the mean squared error of the estimated EQTE $\delta(\tau_n)$ (as defined in (5.1)) as a function of p_u for different tail probability regimes. As expected, the absolute MSE increases as the target tail probability decreases, reflecting the increasing extremeness of the target quantile and the reduced effective sample size available for tail calibration.

The extended range shows clear threshold dependence, particularly for the two most extreme targets. The smallest observed MSE occurs at $p_u = 0.74$ for $(1 - \tau_n) = 1/n$ and at $p_u = 0.80$ for $(1 - \tau_n) = 5/(n \log n)$; in both cases, MSE rises substantially towards $p_u = 0.95$. This pattern reflects the bias–variance trade-off induced by the heavy-tailed design. A high threshold leaves few exceedances for fitting the GPD tail and estimating the extreme treatment effect, increasing variance and destabilising the most extreme extrapolations. Lowering the threshold increases the effective tail sample size and initially reduces variance, but at still lower thresholds the increase in finite-sample bias offsets part of this gain. The less extreme target $(1 - \tau_n) = 5/n$ has a much smaller MSE throughout and a flatter, noisier profile because it requires less extrapolation. Overall, intermediate thresholds provide the most favourable bias–variance balance in this design.

5.5.4.2 Robustness to grid size K

Under the same experimental setting as Section 5.5.4.1, namely the heavy-tailed Pareto design $M_3^{(\mathrm{H})}$ with $n = 1000$, we investigate robustness with respect to the grid size used in the numerical integration detailed in Section 5.3.4. We vary the grid resolution over $\{100, 200, \ldots, 2000\}$ while fixing the intermediate order at $k = n^{0.65}$, equivalently $p_u = 1 - n^{0.65}/n$, and keeping all other components unchanged.

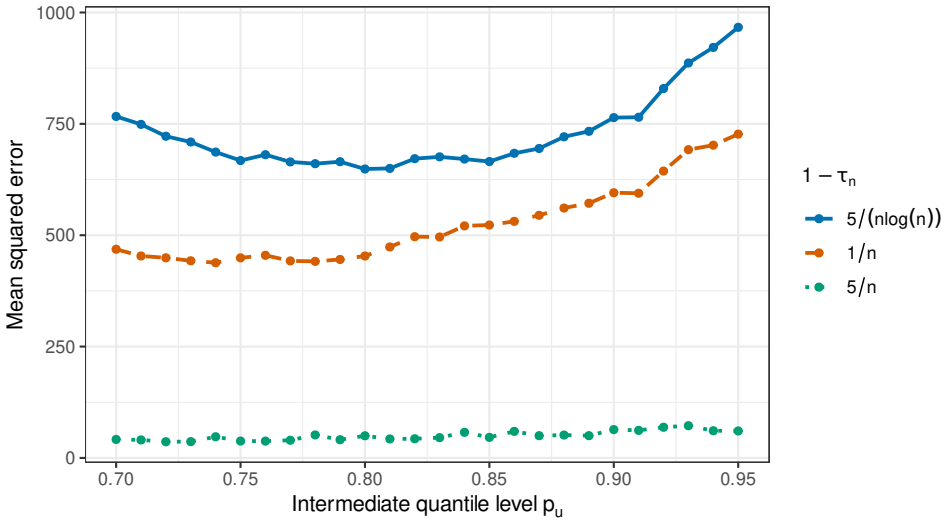

Figure 5.3: Mean squared error of the TIEE–IPW estimator of the EQTE $\delta(\tau_n)$ in (5.1) under the heavy-tailed Pareto design $M_3^{(\mathrm{H})}$ as a function of the intermediate quantile level $p_u \in [0.70, 0.95]$ for $(1-\tau_n) \in \{5/(n\log n), 1/n, 5/n\}$. For replicate estimates $\hat{\delta}_b(\tau_n)$ and data-generating value $\delta_0(\tau_n)$, MSE is $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n)-\delta_0(\tau_n)\}^2$. The experiment uses 1000 Monte Carlo replicates; B denotes the number of successful fits for each setting and ranges from 996 to 1000.

Figure 5.4 examines the robustness of the TIEE-IPW estimator of $\delta(\tau_n)$ to the grid size K under three tail probability regimes. Across the range of grid resolutions considered, the MSE remains stable, with no systematic deterioration as K increases. This indicates that the performance of TIEE is not driven by fine-tuning of the grid resolution and that relatively coarse grids already provide sufficient accuracy. The result suggests that the integrated estimating equation is numerically well behaved and that the proposed implementation is robust to the discretisation choices required for practical computation.

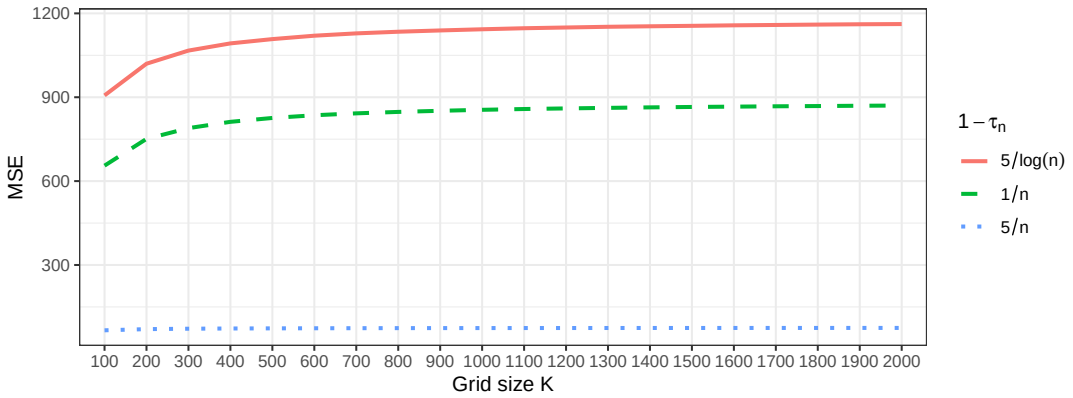

Figure 5.4: Mean squared error (MSE) of the TIEE–IPW estimator of the EQTE $\delta(\tau_n)$ in (5.1) under the heavy-tailed Pareto design $M_3^{(\mathrm{H})}$ as a function of the numerical-integration grid size K for $(1-\tau_n) \in \{5/(n\log n), 1/n, 5/n\}$. For replicate estimates $\hat{\delta}_b(\tau_n)$ and data-generating value $\delta_0(\tau_n)$, MSE is $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n)-\delta_0(\tau_n)\}^2$.

5.5.4.3 Robustness to model misspecification

In the extreme quantile regime, the outcome model is inherently misspecified because the tail behaviour cannot be learnt directly from the observed data. Therefore, the classical notion of double robustness (consistency provided either the propensity score or the outcome model is correctly specified) does not strictly apply here. Nevertheless, our simulation studies in Sections 5.5.2 and 5.5.3 demonstrate that the proposed TIEE estimator exhibits robustness to tail-model misspecification, as the data are generated from non-GPD heavy-tailed distributions whilst the GPD is imposed for extrapolation.

We further investigate robustness to propensity score misspecification by intentionally fitting incorrect propensity score models. Specifically, we consider three misspecification scenarios: (1) omitting the quadratic term in (5.22), (2) using a linear model when the true model is quadratic, and (3) introducing spurious interaction terms. We can see in Table 5.3 that our estimator of the EQTE $\delta(\tau_n)$ continues to deliver small bias and valid inference under these perturbations.

Table 5.3: Bias and MSE of the TIEE–IPW estimator of the EQTE $\delta(\tau_n)$ in (5.1) under different propensity-score specifications. The "True Value" is the EQTE computed from the data-generating distributions by a large-sample Monte Carlo approximation. Writing this value as $\delta_0(\tau_n)$ and the estimate in replicate b as $\hat{\delta}_b(\tau_n)$, Bias and MSE are $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n)-\delta_0(\tau_n)\}$ and $B^{-1}\sum_{b=1}^{B}\{\hat{\delta}_b(\tau_n)-\delta_0(\tau_n)\}^2$, respectively.

$1-\tau_n$	True Value	Propensity Score Model	Bias	MSE
$5/(n\log n)$	45.39	True Model (Quadratic)	-5.438	374.19
		Polynomial Basis	-5.601	397.00
		Misspecified Form (Linear)	-3.903	386.64
		Misspecified Link (Logit)	-3.269	551.88
		Spurious Covariates	-3.472	581.16
$1/n$	38.38	True Model (Quadratic)	-3.837	275.01
		Polynomial Basis	-3.838	283.39
		Misspecified Form (Linear)	-2.516	291.51
		Misspecified Link (Logit)	-2.199	361.22
		Spurious Covariates	-2.103	370.79
$5/n$	16.53	True Model (Quadratic)	-1.338	29.53
		Polynomial Basis	-1.354	29.33
		Misspecified Form (Linear)	-1.009	31.11
		Misspecified Link (Logit)	-1.287	29.74
		Spurious Covariates	-1.285	29.89

Taken together, these findings indicate that the TIEE estimator remains numerically robust when both the tail model and the propensity score model are misspecified. This double robustness in the practical sense (though not in the formal asymptotic sense) is a valuable property for applied work, where perfect model specification is rarely achievable.

5.6 Data application

5.6.1 Data source, motivation and confounding effects

Extreme event attribution (EEA) studies have traditionally relied on climate model ensembles to contrast a factual world under anthropogenic forcing with a counterfactual world in which such forcing is removed. Within this model-based literature, attribution is commonly framed either probabilistically, through changes in the likelihood of exceeding a prescribed threshold (van Oldenborgh et al., 2021), or within a storyline framework, by assessing thermodynamic intensification conditional on observed or prescribed atmospheric circulation patterns (Shepherd, 2016). While powerful, both approaches generally frame attribution in terms of exceedance probabilities or conditional thermodynamic responses, and do not directly target changes in extreme quantiles or return levels.

In this application, we adopt a causal, observational attribution framework that is complementary to both probabilistic and storyline approaches. Rather than targeting changes in exceedance probabilities, we focus on changes in the extreme tail of the outcome distribution, quantified through an extreme quantile treatment effect. At the same time, by conditioning on large-scale atmospheric circulation, our approach aligns with the physical logic underpinning storyline attribution, while formalising it within a causal counterfactual framework that allows treatment effects to be estimated directly from historical observations, without reliance on climate model structure.

We apply the proposed TIEE framework to the EEAR-Clim dataset (Bongiovanni et al., 2025), a high-density observational dataset of daily precipitation (in mm) and air temperature covering the Extended European Alpine Region. The outcome $\boldsymbol{Y}$ is the seasonal 7-day maximum precipitation in the Austrian Alps (Tirol), a region characterised by strong orographic gradients and heavy-tailed precipitation distributions. Defining the treatment $\boldsymbol{D}$ as the modern climate era (1995-2020) versus a historical baseline (1950-

1980), we target the extreme quantile treatment effect at level $\tau = 1 - 0.01 = 0.99$, which under stationary assumptions corresponds to the 100-year return period. This estimand will address the attribution question of how much more intense are very rare precipitation extremes under recent warming, conditional on comparable atmospheric circulation states.

There are two major challenges for observational causal attribution of extremes: severe data sparsity in the tail and the presence of complex meteorological confounders, particularly atmospheric circulation. Valid attribution then requires separating the thermodynamic influence of anthropogenic warming from variability associated with large-scale atmospheric circulation (Shepherd, 2016). This distinction underpins both probabilistic and storyline approaches to extreme event attribution, where circulation is either sampled across ensembles or explicitly conditioned upon to assess thermodynamic intensification under a given dynamical setup. In the present observational causal framework, this separation is formalised by treating atmospheric circulation as a confounding factor, since circulation affects both the effective climate state (historical versus modern) and the occurrence of extreme precipitation events. Accordingly, valid causal attribution in our setting requires conditioning on, adjusting for, or otherwise accounting for circulation variability in order to isolate the thermodynamic contribution of anthropogenic forcing to extreme precipitation. To address this, we construct a confounding set $\boldsymbol{X}$ derived from the ERA5 reanalysis data (ECMWF; Hersbach et al. 2020[1]), specifically targeting the dynamical state of the atmosphere. The selected covariates include:

1. **Synoptic-scale circulation fields**: ERA5 Single Level sea-level pressure (SLP), characterising surface pressure systems; geopotential height at 500 hPa (Z_{500}), describing mid-tropospheric circulation; and wind intensity at 850 hPa (Wind_{850}), capturing lower-tropospheric flow.
2. **Large-scale teleconnection indices**: The North Atlantic Oscillation (NAO) and the Arctic Oscillation (AO) indices, which drive the primary modes of climate variability in the European region.

1. **Data**: Copernicus Climate Change Service (C3S) (2023). ERA5 hourly data on single levels [DOI: 10.24381/cds.adbb2d47] and pressure levels [DOI: 10.24381/cds.bd0915c6] from 1940 to present. Copernicus Climate Data Store. Accessed Sept. 2025.

By conditioning on this set, we implement a dynamical adjustment strategy that aims to separate the thermodynamic effects of anthropogenic warming from variability associated with large-scale atmospheric circulation. We explicitly exclude thermodynamic state variables, such as local temperature, as these lie on the causal pathway through which climate change affects precipitation extremes; conditioning on them would block the very effect of interest and lead to over-adjustment. We also exclude slowly varying oceanic indices, such as the Atlantic Multidecadal Oscillation (AMO) and the Pacific Decadal Oscillation (PDO), as their strong multi-decadal persistence would substantially reduce overlap between historical and modern periods, potentially violating the positivity assumption (Assumption 5.1(4)). The propensity score $\pi(\boldsymbol{X}) = \Pr(\boldsymbol{D} = 1|\boldsymbol{X})$, where $\boldsymbol{D} = 1$ denotes the modern climate period, is estimated using the following logistic regression model:

$$\log\left(\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right) = \beta_0 + \sum_{r=1}^{2} \alpha_r Z_{500}^r + \sum_{r=1}^{2} \gamma_r \mathrm{SLP}^r + \delta \mathrm{Wind}_{850} + \eta \mathrm{NAO} + \sum_{r=1}^{4} \lambda_k \mathrm{AO}^r.$$

Here, we include quadratic terms for geopotential height (Z_{500}) and sea-level pressure (SLP), linear terms for wind intensity and the NAO, and a fourth-degree polynomial for the AO. This flexible specification ensures that the estimated weights effectively balance large-scale circulation patterns between the two periods. Thus, by controlling for Z_{500}, SLP, wind intensity, and teleconnection indices, we isolate the thermodynamic contribution of anthropogenic forcing from purely dynamical circulation variability.

5.6.2 Results

We applied the TIEE estimator, alongside the Causal Hill and Zhang-Firpo estimators, to the 23 meteorological stations listed in the first column of Table 5.4. Table 5.4 addresses the attribution question posed in Section 5.6.1 by quantifying changes in the intensity of very rare precipitation events (99% quantile) under recent warming, conditional on comparable atmospheric circulation states. The estimates vary across methods and stations. TIEE produces positive and statistically significant estimates at several stations, whereas the unadjusted empirical and alternative causal estimates are more mixed.

At Landeck, TIEE estimates a positive effect of 2.97 mm with a 90% confidence interval excluding zero, whereas Zhang–Firpo estimates −0.20 mm with an interval containing zero. TIEE also produces narrower intervals at several stations; at Bruckmur, its interval is about 4–5 mm wide compared with about 18 mm for Zhang–Firpo. At both Landeck and Bruckmur, the Causal Hill intervals include zero while the TIEE intervals do not.

Across the 23 sites, the mean estimated effects are approximately 2.18 mm for TIEE, 1.75 mm for Zhang–Firpo, and 1.78 mm for Causal Hill. The TIEE estimates are also less dispersed across stations.

Table 5.4: Estimates of the 99% extreme quantile treatment effect $\delta(0.99) = Q_{Y(1)}(0.99) - Q_{Y(0)}(0.99)$ from (5.1), in millimetres, for 23 Austrian stations. Columns report the unadjusted empirical contrast and the Causal Hill, Zhang–Firpo, and TIEE–IPW estimates. Bracketed values are 90% confidence intervals; bold entries indicate statistical significance at the 10% level, meaning that the interval excludes zero.

Station	Empirical	Causal Hill	Zhang-Firpo	TIEE
Aspang	-4.92	-1.81 [-6.53, 2.90]	-3.70 [-10.05, 4.73]	-1.25 [-3.60, 1.11]
Bruckmur	5.65	1.95 [-1.98, 5.89]	4.20 [-5.07, 12.70]	**3.01 [0.92, 5.09]**
Deutschlandsberg	-2.67	-4.07 [-8.31, 0.17]	-4.90 [-14.91, 3.26]	**-3.69 [-6.34, -1.04]**
Doellach	5.22	2.85 [-0.07, 5.78]	4.30 [-0.86, 11.16]	**3.60 [1.90, 5.30]**
Feuerkogel	0.82	6.13 [-0.15, 12.42]	**7.90 [2.64, 14.76]**	1.30 [-1.89, 4.49]
Freistadt	3.21	**5.04 [0.51, 9.57]**	4.30 [-1.51, 10.27]	**5.41 [2.84, 7.98]**
Graz Universitaet	-1.62	-2.34 [-6.33, 1.66]	-3.10 [-10.90, 8.90]	-1.53 [-3.89, 0.82]
Hohenau	-0.25	2.32 [-1.21, 5.85]	1.90 [-4.82, 6.78]	0.47 [-1.41, 2.36]
Kremsmuenster Tawes	-3.84	-2.05 [-6.58, 2.48]	-0.50 [-4.95, 3.42]	**-3.61 [-6.05, -1.17]**
Landeck	1.45	0.94 [-1.45, 3.33]	-0.20 [-2.72, 2.93]	**2.97 [1.52, 4.43]**
Mariazellst. Sebastian Flugfeld	-0.49	2.39 [-2.58, 7.36]	0.10 [-3.42, 2.40]	1.24 [-1.15, 3.63]
Mayrhofen	1.34	0.44 [-2.60, 3.48]	1.20 [-6.11, 8.63]	0.68 [-0.94, 2.30]
Muerzzuschlag	9.24	**8.55 [4.86, 12.24]**	**6.70 [0.51, 11.46]**	**8.99 [7.00, 10.99]**
Patscherkofel	3.11	1.61 [-1.32, 4.54]	2.50 [-0.02, 5.86]	**1.96 [0.33, 3.60]**
Puchbergschneeberg	6.27	**8.18 [3.66, 12.70]**	8.30 [-1.06, 16.81]	**7.25 [4.89, 9.62]**
Retzwindmuehle	6.88	**6.17 [3.12, 9.22]**	**7.00 [4.14, 9.23]**	**5.43 [3.80, 7.07]**
Reutte	4.51	**8.94 [3.49, 14.40]**	9.70 [-7.58, 22.47]	**5.46 [2.64, 8.29]**
Schoppernau	7.25	**8.42 [3.04, 13.81]**	**11.00 [0.28, 21.58]**	**5.41 [2.86, 7.96]**
Schroecken	3.05	4.00 [-0.91, 8.90]	4.00 [-4.87, 20.37]	**3.08 [0.47, 5.68]**
St. Jakobdef.	0.41	-2.32 [-5.74, 1.11]	-0.40 [-6.15, 3.68]	-0.24 [-2.15, 1.67]
St. Poeltenlandhaus	4.23	3.87 [-1.33, 9.07]	5.20 [-3.28, 12.88]	1.65 [-0.92, 4.21]
Stift Zwettl	4.45	3.44 [-0.53, 7.41]	**5.30 [1.74, 9.44]**	**2.24 [0.09, 4.39]**
Waizenkirchen	-4.84	-3.10 [-6.83, 0.63]	-1.30 [-7.78, 4.26]	**-4.88 [-6.90, -2.87]**

5.7 Conclusion

Estimating causal treatment effects at extreme quantiles is a critical task in fields such as climate science and econometrics, yet it poses significant challenges for standard causal inference methods. Traditional approaches, including standard quantile treatment effect estimators and the inverting estimating equation framework, are ill-suited for the extreme tails due to density degeneracy and asymptotic failure as quantile levels approach 0 or 1.

This chapter introduced the Tail-Calibrated Inverse Estimating Equation framework, a principled approach that contributes to bridging the gap between causal inference and Extreme Value Theory. By embedding the GPD model within an integrated estimating equation, this framework smooths the tail estimation process. It targets the quantile directly through a globally solved moment condition, rather than relying on the sequential estimate-then-invert procedure of plug-in methods, thereby offering superior numerical stability and variance reduction as demonstrated in Sections 5.5 and 5.6. Note that although our implementation relies on the GPD for tail calibration, the proposed TIEE framework is not intrinsically tied to this specific choice. At a conceptual level, TIEE requires a tail model that delivers uniformly consistent estimation of tail probabilities or quantiles beyond a high threshold, rather than adherence to a particular parametric family. Under Assumption 5.4, which places the potential outcome distributions in a maximum domain of attraction, exceedances above a sufficiently high threshold admit a GPD limit, making the GPD a natural and theoretically justified default. However, this assumption is semi-parametric in nature and governs only the rate of tail decay. Consequently, the GPD is introduced into the framework as a convenient model-assisted representation of the tail, rather than as a structural requirement. Alternative tail specifications, including Weibull-type tails, extended GPD formulations, or other tail models satisfying the usual uniform consistency condition, can be substituted without altering the form of the integrated estimating equation or the causal identification strategy. This modularity underscores the generality of the TIEE approach and highlights its flexibility in accommodating different tail modelling choices, while preserving its ability to deliver stable inference for extreme quantile treatment effects.

We established the key theoretical properties of the TIEE estimator, proving its identification, consistency, and asymptotic normality, thereby ensuring valid statistical inference in the extreme-tail regime. Our extensive simulation studies corroborated these theoretical findings. The TIEE estimator (specifically TIEE-IPW) demonstrated superior performance, yielding substantially lower MSEs compared to existing state-of-the-art estimators across a wide range of heavy-tailed and light-tailed scenarios. Notably, our analysis in Section 5.5.4.3 also revealed that the TIEE-IPW estimator is highly robust to misspecification of the propensity score model, a highly valuable property for practical applications where the true model is unknown.

It is important to notice that further improvements on the propensity score estimation could be achieved using flexible non-parametric or machine learning approaches, which can in some cases better accommodate complex confounding structures. Methods such as generalised additive models, random forests, and gradient boosting machines are data-driven and naturally capture non-linearities and interactions without requiring *a priori* specification of the functional form. Leveraging these modern estimators for $\pi(x)$ reduces the risk of model misspecification and thus improves the practical robustness of the TIEE estimator. Our procedure is fully compatible with flexible nuisance function estimation and can incorporate such techniques seamlessly.

The practical utility of our framework was demonstrated in an application to observational precipitation records from the EEAR-Clim dataset. By comparing 7-day maximum precipitation in Austria between a historical (1950-1980) and a modern (1995-2020) climate period while controlling for atmospheric confounders, we illustrated how TIEE provides a data-driven, robust alternative to climate model ensembles for extreme event attribution.

Future work could focus on extending this framework in several directions. First, to accommodate high-dimensional confounders, potentially by integrating the flexible machine learning estimators discussed above, and second, to handle multivariate or spatial extreme events, moving beyond the current univariate outcome setting. Additionally, the proposed methodology extends naturally to continuous treatments. In this setting, the binary treatment signal function can be replaced by a continuous treatment analogue based on the generalised propensity score or related doubly robust constructions (Hirano and Imbens,

2004; Kennedy et al., 2017), yielding an unbiased estimating function for the distribution function of the potential outcome at a given treatment level. Once such a signal is specified, the integrated estimating equation and tail calibration steps of the TIEE framework remain unchanged. As a result, extreme quantile response functions or contrasts across treatment levels can be estimated without introducing additional assumptions beyond those standardly required for causal identification with continuous treatments (Hirano and Imbens, 2004). Finally, while this work focuses on climate extremes, the TIEE framework addresses a universal challenge: causal inference in the sparse tails of confounded distributions. This methodology offers potential applications in quantitative finance, for isolating the impact of policy interventions on systemic risk measures like Value-at-Risk, and in epidemiology, for quantifying intervention effects on peak infection rates where average-based metrics often mask the critical tail behaviour. These extensions underscore the broad utility of tail-calibrated inference beyond environmental sciences.

Chapter 6

Causal discovery in extremes

This chapter develops a causal discovery framework for multivariate extremes. It introduces tail-induced asymmetry to orient edges and recover sparse propagation structures in heavy-tailed systems, with applications to hydrological and financial risk networks.

A version of this chapter has been submitted for publication under the title *Causal Discovery in Multivariate Extremes via Tail Asymmetry*. Preprint available at arXiv:2604.21620.

6.1 Introduction

Extreme events often propagate through interconnected systems, making directional inference central to risk assessment and intervention. In environmental, financial, and infrastructural settings, the key question is not merely whether variables co-exceed in the tail, but whether an extreme at one variable helps generate an extreme at another (Asadi et al., 2015; Gissibl et al., 2018). This question is especially difficult in the tail regime, where extreme observations are sparse, tail dependence can differ sharply from bulk behaviour, and latent common drivers can simultaneously push many variables into the tail, inducing strong and largely symmetric co-exceedance that swamps the directional signatures of local propagation and makes both skeleton screening and edge orientation unreliable without explicit adjustment (Resnick, 2007). Classical causal discovery procedures reviewed in Section 3.4 such as the PC algorithm (Spirtes et al., 2000), LiNGAM (Linear Non-Gaussian Acyclic Model; Shimizu et al. 2006), and NOTEARS (Zheng et al., 2018) are therefore poorly matched to this setting. Their inferential logic is calibrated to bulk-level conditional independence, likelihood structure, or non-Gaussian variation rather than to tail events.

As discussed in the review of extremal graphical and directed models in Section 3.4, extreme-value statistics has led to substantial progress in graphical modelling for tail dependence. Along the *undirected line*, extremal graphical models show that sparse tail dependence can be represented through graphical structure, making extremal graph learning substantially more tractable (Engelke and Hitz, 2020). Subsequent work develops structure learning for extremal tree models and related extensions of this line (Engelke and Volgushev, 2022). Along the *directed line*, approaches have emerged under recursive max-linear or related heavy-tailed structural equation model formulations. Early work formulates Bayesian-network structure for max-linear systems (Klüppelberg and Lauritzen, 2019). Subsequent contributions study identifiability, estimation, and causal discovery in recursive max-linear or heavy-tailed models (Gissibl et al., 2021; Gnecco et al., 2021b; Pasche et al., 2023). More recently, Jiang et al. (2025) introduced separation-based causal discovery for extremes, showing that transformed-linear structural causal models can support constraint-based DAG learning through partial tail uncorrelatedness.

Taken together, these developments show that directional learning in extremes is possible, but not yet in the form needed for fully data-driven causal discovery in larger multivariate systems. The undirected line yields sparse tail structure without causal orientation, whereas the directed line typically relies on stronger structural restrictions, narrower graph classes, or constraint-based separation logic. The present chapter takes a complementary route, focusing on score-based orientation through tail-prediction asymmetry and on proxy adjustment for latent common shocks.

We address this gap through a structural insight. Extreme-value processes driven by regularly varying innovations obey the single big jump principle (Resnick, 2007), under which tail events are typically generated by one dominant shock rather than by the accumulation of many moderate contributions. In recursive extremal systems, this creates a directional imbalance. Conditioning on an extreme at a cause sharply constrains the corresponding effect; conditioning on an extreme at an effect does not identify its source, since the exceedance may have arisen through propagation from another variable or through a local innovation. Predicting forward is therefore systematically easier than predicting backward in the tail regime. We call this phenomenon *tail-induced asymmetry*. Building on this insight, we propose *Sparse Structure Discovery in Multivariate Extremes* (S3ME), a

two-stage framework for causal discovery in multivariate extremes. The first stage screens for a sparse candidate skeleton via proxy-adjusted penalized regression, where the proxy adjustment attenuates symmetric dependence induced by latent common shocks. The second stage orients edges by exploiting tail-induced asymmetry through a score-based procedure. The framework is supported by an identifiability result for the canonical bivariate case, a high-dimensional sure-screening guarantee for the skeleton stage and an oracle consistency result for a direct-fit global orientation score, and a robustness analysis showing that proxy adjustment suppresses confounding-induced false positives where existing directed methods deteriorate. We apply the method to 31 gauging stations locations in the upper Danube river network and a tail-risk network of 103 S&P 500 constituents spanning all eleven GICS sectors, settings substantially larger than those in existing directed extremal work. Our method recovers sparse, interpretable propagation structures without any prior graph specification.

This chapter makes three contributions. First, we identify tail-induced asymmetry as an exploitable source of causal directionality in heavy-tailed systems, a mechanism that has not previously been exploited for graph orientation in high-dimensional, fully data-driven settings. Second, we show that this asymmetry supports oracle-consistent edge orientation under a direct-fit global score in high-dimensional settings without a prespecified graph, extending the reach of causal discovery in extremes well beyond existing methods. Third, we develop a proxy-adjustment procedure that remains stable under latent confounding at a scale where competing directed approaches accumulate large numbers of spurious edges.

The remainder of the chapter is organized as follows. Section 6.2 establishes the core identifiability result based on tail-induced asymmetry. Section 6.3 presents the two-stage methodology. Section 6.4 provides theoretical guarantees for high-dimensional sure screening, oracle direct-fit orientation consistency, and robustness to latent confounding. Section 6.5 evaluates finite-sample performance through simulations, Section 6.6 applies the method to hydrological and financial data and Section 6.7 concludes the chapter. Code to reproduce the simulation studies and real-data applications in this chapter is available at https://github.com/MengranLi-git/S3ME.

6.2 Identifiability via tail asymmetry

6.2.1 Recursive extremal SEMs and tail prediction risk

In complex systems, tail events may propagate along directed pathways. To represent this mechanism, we consider a recursive extremal structural equation model. The graphical notation follows the SCM/DAG framework introduced in Section 2.1.3.

Assumption 6.1 (Recursive Extremal SEM). Let $X = (X_1, \dots, X_p)^\top$ be a strictly positive random vector. There exists a directed acyclic graph (DAG) $\mathcal{G} = (V, E)$ with parent sets $\mathrm{pa}(j)$ such that, for $j = 1, \dots, p$,

$$X_j = f_j\big(\{X_i : i \in \mathrm{pa}(j)\}\big) + \epsilon_j, \tag{6.1}$$

where: (i) $f_j(\cdot)$ is non-decreasing in each argument and homogeneous of order 1, i.e., $f_j(cx) = c f_j(x)$ for all $c > 0$; (ii) $\epsilon_1, \dots, \epsilon_p$ are mutually independent, strictly positive, and regularly varying with common tail index $\alpha > 0$.

Assumption 6.1 includes standard recursive extremal SEMs, including max-linear and extremal additive constructions, and implies multivariate regular variation for X. The single big jump principle then creates a characteristic directional imbalance in the tail, whereby conditioning on an extreme at a parent can substantially reduce uncertainty about its descendants, whereas conditioning on an extreme at a child does not uniquely localize its source (Resnick, 2007, p. 217). We formalize this imbalance through tail prediction risk. Let Z denote the tail-domain representation, whose explicit construction is given in Section 6.3.1. We call a function $h : \mathbb{R} \to \mathbb{R}$ a max-linear envelope predictor if it has the form $h(z) = \max(z, c) + d$ for some $c, d \in \mathbb{R}$, and write $\mathcal{H}$ for the class of all such predictors.

Definition 6.1 (Tail prediction risk)**.** *At threshold u, define the tail prediction risk of Z_j from Z_i by*

$$R^\star_{j|i}(u) \;:=\; \inf_{h \in \mathcal{H}} \mathbb{E}\Big[\,|Z_j - h(Z_i)| \,\Big|\, Z_i > 0\Big], \qquad i \neq j. \tag{6.2}$$

For nodewise parent-set analysis, we extend Definition 6.1 as follows. For any nonempty candidate parent set $A \subseteq V \setminus \{j\}$, define the multivariate max-linear envelope class

$$\mathcal{H}_A := \left\{ h_A(z_A) = \max\Big(c_0, \max_{m \in A}(z_m + c_{m\to j})\Big) : \; c_0 \in \mathbb{R}, \; c_{m\to j} \in \mathbb{R} \right\},$$

and the corresponding population tail prediction risk

$$R^{\star}_{j|A}(u) := \inf_{h_A \in \mathcal{H}_A} \mathbb{E}\left[|Z_j - h_A(Z_A)| \,\middle|\, \max_{m\in A} Z_m > 0 \right].$$

For $A = \{i\}$, this reduces (up to reparametrization) to the bivariate definition above. We say that (i, j) exhibits *tail-induced asymmetry* at level u if $R^{\star}_{j|i}(u) < R^{\star}_{i|j}(u)$. Intuitively, when a causal relationship $i \to j$ exists, predicting j from i is easier than the reverse. A large shock at i reliably drives j into the tail, so Z_i is a tight predictor of Z_j. Conversely, backward prediction is more ambiguous. An extreme at j may have been driven by a shock at i or may have arisen from j's own innovation ϵ_j, and given only Z_j, the two are indistinguishable. The prediction risk therefore satisfies $R^{\star}_{j|i} < R^{\star}_{i|j}$, and this asymmetry reveals the causal direction in a way that standard tail dependence measures, which are symmetric in i and j, cannot.

6.2.2 Asymptotic identifiability

While Assumption 6.1 allows for general regularly varying innovations with index $\alpha > 0$, the directional tail signatures we study are invariant under monotone marginal transformations. For the theoretical analysis that follows (Theorem 6.1 and Proposition C.2 in Appendix C.4), it is standard practice to work with the canonical representation. Therefore, without loss of generality, we assume the innovations have been standardized to unit Fréchet scale ($\alpha = 1$) via the probability integral transform, so that $F_{\epsilon_j}(x) = \exp(-1/x)$ for each j. This specification provides a standard heavy-tailed benchmark in which the asymmetry mechanism can be analyzed in closed form, and it serves as the basic identifiability template for the orientation step developed later.

Theorem 6.1 (Bivariate identifiability via tail asymmetry). *Let* $X = (X_1, X_2)^\top$ *follow the bivariate recursive max-linear model* $X_1 = \epsilon_1$ *and* $X_2 = \max(cX_1, \epsilon_2)$*, where* ϵ_1, ϵ_2 *are independent Fréchet*(1) *innovations and* $c > 0$*. Let* Y_j *denote the exact unit-Pareto standardization of* X_j*, and let* $Z_j = \log Y_j - \log u$ *denote the corresponding log-tail coordinates. Then the forward risk converges to zero, i.e.,* $R^\star_{2|1}(u) \to 0$*, as* $u \to \infty$*, whereas the reverse risk is strictly bounded away from zero:*

$$\liminf_{u\to\infty} R^\star_{1|2}(u) \geq \frac{\log 2}{c+1} > 0. \tag{6.3}$$

Consequently, any monotone ℓ_1*-based scoring criterion evaluated at a sufficiently large threshold* u *strictly prefers the true direction* $1 \to 2$ *over* $2 \to 1$*, yielding asymptotic identifiability under max-linear envelope prediction.*

The two parts of Theorem 6.1 (whose proof is deferred to Appendix C.2) reflect the asymmetry described above. In the forward direction, when X_1 is very large, $X_2 = \max(cX_1, \epsilon_2) \approx cX_1$ dominates, so $Z_2 \approx Z_1 + \log c$ and the prediction error vanishes. In the reverse direction, an extreme X_2 could have arisen from a large X_1 or from a large ϵ_2, and the two scenarios imply very different values of X_1; this residual ambiguity is what keeps $R^\star_{1|2}$ bounded away from zero. The lower bound $\log 2/(c+1)$ decreases as c increases, since a stronger causal link leaves a larger footprint on X_2, making the source somewhat easier to recover, but the ambiguity never fully disappears at any finite c.

6.2.3 Nodewise and global identifiability

Theorem 6.1 establishes the directional mechanism in the canonical bivariate case. We now extend this to the multivariate setting. The key challenge is that each node may have multiple candidate parents, so identifiability must simultaneously address two questions: underfitting (omitting a true parent) and overfitting (including spurious variables). These are controlled by two assumptions.

Assumption 6.2 (Non-vanishing conditional prediction error for omitted parents). Let $\mathcal{G}^* = (V, E^*)$ be the true DAG, fix a node $j \in V$, let $\mathrm{pa}^*(j)$ denote its true parent set, and let $\mathrm{nb}^*(j) = \{i : \{i, j\} \in E^*_{\mathrm{skel}}\}$ denote the neighbors of j in the true undirected skeleton. For any candidate parent set $A \subseteq \mathrm{nb}^*(j)$ with $A \not\supseteq \mathrm{pa}^*(j)$, the omission of at least one true parent induces a non-vanishing increase in population prediction error in the tail. Specifically, the optimal ℓ_1 tail prediction risk for node j given parent set A satisfies $\liminf_{u\to\infty}\left[R^\star_{j|A}(u) - R^\star_{j|\mathrm{pa}^*(j)}(u)\right] > 0$.

This assumption abstracts the multivariate extension of the bivariate tail asymmetry established in Theorem 6.1: conditioning on a strict subset of the true parents does not eliminate the directional signal from the missing parent.

To operationalize the tail prediction risk for graph learning, we fix a sufficiently large threshold u where the asymptotic tail asymmetry of Assumption 6.2 takes effect and work directly with $R^\star_{j|A}(u)$ as the population target. Thus, at the population level, we do not include sample-size-dependent complexity penalties; those enter only in the empirical score used for finite-sample orientation in Section 6.3.2.

Assumption 6.3 (No population risk gain from spurious supersets). Under the same notation, if $A \supseteq \mathrm{pa}^*(j)$, then adding non-parent variables does not improve the population tail prediction risk $R^*_{j|A}(u) \geq R^*_{j|\mathrm{pa}^*(j)}(u)$ for sufficiently large u. In the max-linear benchmark, Proposition C.1 (Appendix C.3) verifies the sharper equality case.

Assumption 6.2 controls underfitting, ensuring that omitting a true parent incurs a strictly positive population risk increase. Assumption 6.3 controls overfitting, ensuring that adding spurious parents cannot improve population risk. Together they imply that the true parent set is the unique inclusion-minimal population minimizer at each node.

Theorem 6.2 (Nodewise identifiability). *Let $X = (X_1, \ldots, X_p)^\top$ follow a recursive max-linear SEM on a true DAG $\mathcal{G}^* = (V, E^*)$, and let E^*_{skel} denote the corresponding undirected skeleton. Fix a node $j \in V$. Under Assumptions 6.2 and 6.3, for any fixed and sufficiently large threshold u, the true parent set $\mathrm{pa}^*(j)$ is the unique inclusion-minimal minimizer of $R^*_{j|A}(u)$ over all skeleton-compatible candidate parent sets $A \subseteq \mathrm{nb}^*(j)$.*

Corollary 6.1 (Global identifiability by risk decomposability). *Let*

$$R^*(G; u) = \sum_{j=1}^{p} R^*_{j|\mathrm{pa}_G(j)}(u)$$

be the population decomposable risk over DAGs compatible with the true skeleton. If Theorem 6.2 holds for every node $j \in V$, then $\mathcal{G}^*$ is the unique inclusion-minimal minimizer of $R^*(G;u)$ over the skeleton-compatible class.

The formal proofs are deferred to Appendix C.3. Proposition C.1 (Appendix C.3) verifies the superset non-improvement property in the max-linear subclass; extending this verification to broader homogeneous non-decreasing SEMs remains open.

6.3 Methodology

6.3.1 Skeleton recovery in the tail domain

Standard max-linear specifications satisfying Assumption 6.1 induce multivariate regular variation on X (Klüppelberg and Lauritzen, 2019), placing it in the maximum domain of attraction of a max-stable distribution. We standardize each margin to unit Pareto scale via the probability integral transform $Y_j = \frac{1}{1-F_j(X_j)}$, where F_j denotes the marginal distribution function of X_j. While Section 6.2 works with unit Fréchet innovations for theoretical tractability, the unit Pareto standardization here follows the extremal graphical modeling convention of Engelke and Hitz (2020); the two scales are tail-equivalent (both regularly varying with index 1), and the log-transformation in (6.4) yields approximately exponential tail coordinates that are convenient for subsequent estimation. We then focus on the tail region defined by the exceedance event $\{\|Y\|_\infty > u\}$, where $\|Y\|_\infty = \max_{1\le j\le p} Y_j$, for a sufficiently large threshold u, ensuring that at least one component is extreme; this is the standard conditioning event in the multivariate regular variation framework. Following the extremal graphical modeling framework of Engelke and Hitz (2020), we introduce the log-transformed tail coordinates

$$Z = \log Y - \log u, \tag{6.4}$$

which provide a convenient representation for characterizing extremal conditional dependence in high dimensions. Note that $\{\|Y\|_\infty > u\}$ is equivalent to $\{Z \in \mathcal{L}\}$ with $\mathcal{L} = \cup_{k=1}^{p}\{z : z_k > 0\}$.

To make orientation feasible in high dimensions, we first estimate a sparse undirected candidate skeleton. This stage serves a screening role analogous to sure independence screening in high-dimensional regression, with the objective of reducing the edge set to a manageable superset that retains all true edges with high probability, not to recover the skeleton exactly or to estimate regression coefficients precisely. Removing implausible pairs before orientation keeps the subsequent direction search tractable, while Section 6.4.1 establishes conditions under which the screened skeleton retains all true edges, so that the true DAG remains reachable. For this purpose, we adopt the Hüsler–Reiss model as a working model for the tail law of Z, following Engelke and Hitz (2020).

Assumption 6.4 (Hüsler–Reiss tail approximation). The law of Z restricted to $\mathcal{L}$ is approximated by a Hüsler–Reiss extremal model parameterized by a variogram matrix $\Gamma \in \mathbb{R}^{p\times p}$. Through its associated precision structure Θ, the model admits a Gaussian graphical representation in the sense that extremal conditional independence among the components of Z corresponds to zero entries of Θ (Engelke and Hitz, 2020).

In practice, the Hüsler–Reiss approximation is applied after partialling out a proxy P for latent common shocks, that is, unobserved variables that simultaneously drive multiple nodes into the tail, so that edges in the estimated graph reflect direct tail connections between variables rather than spurious co-exceedance driven by such shocks. Under Assumption 6.4, the conditional distribution of Z_j given Z_{-j} in the tail admits a regression representation whose coefficients are determined by Θ (Engelke and Hitz, 2020); zeros in Θ thus correspond to zero nodewise regression coefficients, which motivates ℓ_1-penalized neighbourhood selection in the spirit of Meinshausen and Bühlmann (2006).

When latent common shocks are present, direct application of nodewise ℓ_1-penalized regression can produce dense candidate graphs, because such shocks induce strong and largely symmetric tail co-exceedance across many pairs of nodes. In such settings, a low-dimensional proxy P for latent common shocks can be included as an unpenalized control in each nodewise regression; as shown in Section 6.5, this adjustment substantially improves recovery when confounding is present while leaving performance essentially unchanged when it is not. When a proxy is used, let $\{(Z^{(i)}, P^{(i)})\}_{i=1}^{k}$ denote the log-exceedances and corresponding proxy values. The regression representation of the Hüsler–Reiss model (Assumption 6.4) implies that Z_j is approximately linear in Z_{-j} in

the tail, so squared-error loss is a natural fit. Sparsity is enforced by an ℓ_1 penalty on the neighbour coefficients β, following the neighbourhood-selection principle of Meinshausen and Bühlmann (2006). The proxy P and intercept α enter without a penalty, so that common-shock variation is always absorbed regardless of the regularisation strength. For each $j \in V$, we estimate

$$(\hat{\beta}^{(j)}, \hat{\gamma}_j, \hat{\alpha}_j) = \arg \min_{\beta \in \mathbb{R}^{p-1}, \gamma \in \mathbb{R}^d, \alpha \in \mathbb{R}} \left\{ \frac{1}{2k} \sum_{i=1}^{k} \left(Z_j^{(i)} - \alpha - \sum_{m \neq j} \beta_{jm} Z_m^{(i)} - \gamma^\top P^{(i)} \right)^2 + \lambda_j \|\beta\|_1 \right\}, \tag{6.5}$$

so that the proxy term and intercept are not penalized while sparsity is enforced only on the cross-node coefficients. The skeleton estimate $\hat{E}_{\text{skel}}$ is constructed by thresholding, retaining predictor m for node j if $|\hat{\beta}_m^{(j)}| > \tau$ for a threshold $\tau \propto \lambda$, and taking the union of retained edges across all nodes.

6.3.2 Score-based edge orientation

Given $\hat{E}_{\text{skel}}$, we orient edges by minimizing a decomposable nodewise score over DAGs compatible with the candidate graph. The score is designed to exploit the asymmetry identified in Section 6.2.2, so that directions that better capture extremal propagation yield smaller tail-domain prediction error.

For a candidate parent set $\text{pa}(j)$ (and proxy P if used), we represent node j by a max-linear envelope in the log-tail domain. For each candidate parent $k \in \text{pa}(j)$, the transmission offset $c_{k \to j}$ (i.e., the log-scale shift in tail level as a shock propagates from k to j) is estimated by a low empirical quantile: $c_{k \to j} = \hat{Q}_q(Z_{\cdot j} - Z_{\cdot k})$ for $q \in (0, 1)$, and the baseline intercept c_0 is defined analogously. For the proxy, each component $r = 1, \ldots, d$ of P has its own offset $c_{r \to j} = \hat{Q}_q(Z_{\cdot j} - P_{\cdot r})$. The fitted envelope for node j at observation t is then

$$\hat{Z}_{tj} = \max \left\{ c_0, \max_{\ell \in \text{pa}(j)} \left(Z_{t\ell} + c_{\ell \to j} \right), \max_{r=1,\ldots,d} \left(P_{tr} + c_{r \to j} \right) \right\}, \tag{6.6}$$

where $\max_{\ell \in \emptyset}(\cdot) := -\infty$ by convention, so the parent term is simply absent when $\text{pa}(j) = \emptyset$. The envelope captures the dominant upstream contribution, with the proxy term absorbing common forcing. The low quantile q provides a stable componentwise calibration of the lower envelope and avoids a separate nonsmooth joint optimisation for every can-

didate parent set. A direct fit would instead choose all offsets appearing in (6.6) jointly to minimise the SAE criterion below. Because the maximum couples these offsets, the componentwise quantile rule is not, in general, the joint SAE minimiser; it is the computational score used by Algorithm 6.1.

For each node j, we evaluate a candidate parent set by the sum of absolute envelope (SAE) residuals under this quantile-calibrated envelope:

$$\mathrm{SAE}_j = \sum_{t=1}^{k} \left| Z_{tj} - \hat{Z}_{tj} \right|, \tag{6.7}$$

and combine this fit measure with an Extended Bayesian Information Criterion (EBIC; Foygel and Drton, 2010) complexity penalty. The local score is defined by

$$\mathrm{Score}(j, \mathrm{pa}(j)) = \frac{k}{2} \log\left(\frac{\mathrm{SAE}_j}{k}\right) + \frac{1}{2}\left(\log k + 2\gamma_{\mathrm{EBIC}} \log p\right) |\mathrm{pa}(j)|, \tag{6.8}$$

where p is the number of observed variables, $\mathrm{SAE}_j > 0$ by construction (since the max-linear envelope predictor generically differs from Z_{tj} on a positive-measure set of tail observations), and the proxy P is excluded from the parent-set penalty. Because the score decomposes across nodes, we optimize the global score $\sum_{j=1}^{p} \mathrm{Score}(j, \mathrm{pa}(j))$ by greedy search over DAGs compatible with $\hat{E}_{\mathrm{skel}}$. The user-specified bound $d_{\max}$ is the maximum number of parents permitted for any node. It should be large enough to include scientifically plausible parent sets, but small relative to the effective tail sample size k, because only the retained exceedances contribute to the orientation score. Starting from the empty DAG on V, the algorithm considers additions only between node pairs in $\hat{E}_{\mathrm{skel}}$, together with deletions and reversals of currently selected edges. Thus, the estimated skeleton restricts the candidate adjacencies rather than supplying the initial directed graph. At each iteration, the algorithm applies the admissible move that yields the largest score reduction. Moves that create directed cycles or violate the maximum in-degree constraint are excluded. The search terminates when no admissible move improves the score, and the

resulting graph is returned as the estimated DAG. The theoretical analysis in Section 6.4.2 concerns the oracle global minimiser of a direct-fit envelope score that jointly optimises the offsets over $\mathbb{R}$. The quantile-calibrated score and greedy search used in Algorithm 6.1 are computational approximations to that theoretical target.

6.3.3 The proposed algorithm

Algorithm 6.1 summarises the full S3ME procedure. When no proxy is available, the proxy offset terms $c_{r\to j}$ and components $P_r^{(i)}$ $(r = 1, \dots, d)$ are omitted from the offset estimation and envelope computation steps, and the max-linear envelope simplifies to $\hat{Z}_j^{(i)} = \max\left\{c_0,\ \max_{\ell\in \mathrm{pa}(j)}\left(Z_\ell^{(i)} + c_{\ell\to j}\right)\right\}$. Computing the empirical marginal ranks and extracting the tail sample from n observations costs $O(np\log n)$. Let T_{L} bound the effective number of coordinate-descent sweeps used for each nodewise Lasso. Since each regression contains $p-1+d$ predictors, the skeleton stage costs $O\{T_{\mathrm{L}}kp(p+d)\}$, with an additional $O(p^2\log p)$ comparison-sorting bound for degree-capped skeleton construction. Let

$$s_{\max} := \max_{1\le j\le p} |\mathrm{nb}(j; E^*_{\mathrm{skel}})|$$

denote the maximum degree of the true skeleton, which measures the local sparsity relevant to both screening and orientation, and write $M = |\hat{E}_{\mathrm{skel}}|$. Each complete greedy scan considers $O(M)$ candidate additions, deletions, and reversals, with both orientations evaluated for each candidate pair. Using linear-time empirical quantile selection, recomputing the local score for an affected node costs $O\{k(d_{\max}+d)\}$. In the current adjacency-matrix implementation, however, an acyclicity check may cost $O(p^2)$ per candidate, and constructing the candidate masks costs $O(p^2)$ per scan. If the search takes I complete scans, an implementation-aware overall bound is therefore

$$O\Big(np\log n + T_{\mathrm{L}}kp(p+d) + p^2\log p + I\left[p^2 + M\{p^2 + k(d_{\max}+d)\}\right]\Big).$$

On the sure-screening event $M = O(ps_{\max})$, this becomes

$$O\Big(np\log n + T_{\mathrm{L}}kp(p+d) + p^2\log p + I\left[p^2 + ps_{\max}\{p^2 + k(d_{\max}+d)\}\right]\Big).$$

For fixed d and T_{L}, the simpler expression $O\{kp^2 + Ikd_{\max}M\}$ captures only the score-evaluation-dominated terms and is not the full cost of the current implementation. The bound $d_{\max}$ therefore plays a dual role, controlling the per-score computational cost while also acting as a structural regulariser on the estimated DAG. For a DAG G, let

$$\mathcal{S}_{\gamma_{\mathrm{EBIC}}}(G) := \sum_{j=1}^{p} \mathrm{Score}\big(j, \mathrm{pa}_G(j)\big)$$

denote the global score. For an admissible move a applied to the current graph $\hat{G}$, define

$$\Delta\mathcal{S}_{\gamma_{\mathrm{EBIC}}}(a; \hat{G}) := \mathcal{S}_{\gamma_{\mathrm{EBIC}}}(\hat{G} \circ a) - \mathcal{S}_{\gamma_{\mathrm{EBIC}}}(\hat{G}),$$

where $\hat{G} \circ a$ is the DAG obtained after applying a. Thus, a negative value denotes an improvement. Two implementation details are worth noting. First, the marginal standardization $\hat{Y}_j^{(i)} = 1/(1-\hat{F}_j(X_j^{(i)}))$ uses the rescaled empirical CDF $\hat{F}_j(x) = \frac{1}{n+1}\sum_{i'=1}^{n} \mathbb{1}(X_j^{(i')} \le x)$ (pseudo-observations with denominator $n+1$), so that $\hat{Y}_j^{(i)} = (n+1)/(n+1-R_j^{(i)})$ where $R_j^{(i)}$ is the rank of $X_j^{(i)}$; this ensures $\hat{Y}_j^{(i)} < \infty$ for all observations. Second, the estimated skeleton is used only as a candidate-edge mask: the orientation search is initialised from the empty DAG, and additions are restricted to node pairs in $\hat{E}_{\mathrm{skel}}$.

Algorithm 6.1: S3ME: Sparse Structural Causal Discovery in Multivariate Extremes

Require: Observations $\{X^{(i)}\}_{i=1}^n$ with $X^{(i)} \in \mathbb{R}^p$, proxy values $\{P^{(i)}\}_{i=1}^n$ with $P^{(i)} \in \mathbb{R}^d$, tail threshold u, regularisation parameters $\{\lambda_j\}$, coefficient threshold τ, offset quantile $q \in (0,1)$, EBIC parameter γ_{EBIC}, maximum in-degree $d_{\max}$.

Output: Estimated DAG $\hat{G}$.

Tail Preprocessing.

Compute $\hat{Y}_j^{(i)} = 1/(1 - \hat{F}_j(X_j^{(i)}))$ via the empirical rank transform.

Retain the k observations with $\|\hat{Y}^{(i)}\|_\infty > u$, re-index them as $i = 1, \dots, k$, and set $Z^{(i)} = \log \hat{Y}^{(i)} - \log u$; retain the corresponding $\{P^{(i)}\}_{i=1}^k$.

Skeleton Estimation.

for $j = 1, \dots, p$ **do**

  Solve the proxy-adjusted nodewise Lasso with penalty λ_j on β and unpenalised proxy P to obtain $\hat{\beta}^{(j)}$.

end for

Form $\hat{E}_{\mathrm{skel}}$ with edges $\{j, \ell\}$ if $\max(|\hat{\beta}_\ell^{(j)}|, |\hat{\beta}_j^{(\ell)}|) > \tau$.

Score-Based Orientation.

Initialise $\hat{G}$ as the empty DAG on vertex set V; use $\hat{E}_{\mathrm{skel}}$ only to restrict admissible edge additions.

repeat

  for each admissible move a (edge addition, deletion, or reversal in $\hat{G}$, with additions restricted to pairs in $\hat{E}_{\mathrm{skel}}$, preserving acyclicity and in-degree $\le d_{\max}$) **do**

    for each node j whose parent set changes under a **do**

      Estimate offsets $c_{\ell \to j} = \hat{Q}_q\big(\{Z_j^{(i)} - Z_\ell^{(i)}\}_{i=1}^k\big)$ for each $\ell \in \mathrm{pa}(j)$; set $c_{r \to j} = \hat{Q}_q\big(\{Z_j^{(i)} - P_r^{(i)}\}_{i=1}^k\big)$ for each $r = 1, \dots, d$; and $c_0 = \hat{Q}_q\big(\{Z_j^{(i)}\}_{i=1}^k\big)$.

      Compute $\hat{Z}_j^{(i)} = \max\big\{c_0,\ \max_{\ell \in \mathrm{pa}(j)}(Z_\ell^{(i)} + c_{\ell \to j}),\ \max_{r=1,\dots,d}(P_r^{(i)} + c_{r \to j})\big\}$ for each $i = 1, \dots, k$.

      Evaluate $\mathrm{SAE}_j = \sum_{i=1}^k |Z_j^{(i)} - \hat{Z}_j^{(i)}|$ and $\mathrm{Score}_{\gamma_{\mathrm{EBIC}}}(j, \mathrm{pa}(j))$.

    end for

  end for

  Set $a^* = \arg\min_a \Delta\mathcal{S}_{\gamma_{\mathrm{EBIC}}}(a; \hat{G})$.

  if $\Delta\mathcal{S}_{\gamma_{\mathrm{EBIC}}}(a^*; \hat{G}) < 0$ **then**

    Apply a^* to $\hat{G}$.

  end if

until no admissible move reduces the score.

return $\hat{G}$.

6.4 Theoretical properties

Let $\mathcal{G}^* = (V, E^*)$ be the true causal DAG with node set $V = \{1, \dots, p\}$, and let $E^*_{\text{skel}} = \big\{\{i, j\} : (i, j) \in E^* \text{ or } (j, i) \in E^*\big\}$ denote the corresponding undirected skeleton. The theoretical analysis reflects the division of labour between the two stages, where the skeleton procedure is a screening device that retains all true edges with high probability, while the orientation step eliminates spurious edges and recovers the correct directions through the EBIC penalty and tail-induced asymmetry respectively.

6.4.1 Sure screening for the skeleton

We work in a high-dimensional extreme-value regime in which both the dimension p and the effective tail sample size k (the number of exceedances retained for inference) tend to infinity, potentially allowing $p \gg k$. We assume $k = k(n)$ is an intermediate sequence with $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. The sparsity requirement used below is $s_{\max}\sqrt{\log p / k} = o(1)$. The logarithmic scaling $s_{\max} = O(\log p)$ is a transparent sufficient regime: it allows the maximum neighbourhood size to increase slowly with dimension while excluding dense graphs. Under this scaling, the rate condition holds whenever $(\log p)^3/k \to 0$.

Under the Hüsler–Reiss working model (Assumption 6.4), the precision structure Θ introduced in Section 6.3.1 induces a covariance-type matrix $\Sigma = \Theta^{-1}$. As shown by Engelke and Hitz (2020), the nodewise regression coefficients in the tail can be expressed analogously to the Gaussian case as $\beta_{jk} = -\Theta_{jk}/\Theta_{jj}$, so zeros in β encode zeros in Θ.

Assumption 6.5 (Regularity conditions for sure screening). Under the Hüsler–Reiss working model and the proxy-adjusted nodewise regression, assume: (i) $s_{\max}$ is sufficiently small relative to $k/\log p$, (ii) the relevant submatrices of $\Sigma = \Theta^{-1}$ are uniformly well-conditioned, (iii) the minimum non-zero tail regression coefficient is above the screening noise level, and (iv) the nodewise score bias from the log-sum-exp approximation is of order at most λ. Precise constants and rate thresholds are deferred to Appendix C.4.

Assumption 6.5 provides the extremal analogs of the standard high-dimensional conditions used in neighborhood selection (Meinshausen and Bühlmann, 2006; Wainwright, 2009). We do not require the irrepresentability condition needed for exact support recovery (Wainwright, 2009, Condition 1); rather, the conditions above ensure stable local design, detectable tail signals, and controlled approximation bias at effective sample size k.

In the tail regime, the log-sum-exp proxy is a uniformly bounded approximation of the max-linear structural equation; a formal statement and proof are given in Proposition C.2 in Appendix C.4.

Theorem 6.3 (Sure screening for the skeleton)**.** *Suppose Assumptions 6.4 and 6.5 hold, and the tail linearization error is uniformly bounded in the sense of Proposition C.2 (Appendix C.4). Let $\hat{E}(\lambda)$ denote the undirected edge set obtained by the union of selected neighborhoods from nodewise ℓ_1-penalized regression (Section 6.3.1), with the proxy included as an unpenalized regressor if used, using regularization parameter $\lambda > C\sqrt{\log p/k}$ for a sufficiently large constant $C > 0$, as is standard in high-dimensional regression theory. Then there exist constants $c_1, c_2, c_3 > 0$ such that for sufficiently large k and p,*

$$Pr\Big(E^{\star}_{\mathrm{skel}} \subseteq \hat{E}(\lambda)\Big) \;\geq\; 1 - c_1 \exp(-c_2 k\lambda^2), \tag{6.9}$$

and, on this event, the number of selected edges satisfies $|\hat{E}(\lambda)| \leq c_3\, p\, s_{\max}$.

Sure screening, rather than exact support recovery, suffices for the subsequent orientation step. Indeed, the sure screening guarantee ensures that the true DAG G^* belongs to the search space $\mathcal{G}(\hat{E}_{\mathrm{skel}})$, so the correct orientation is always reachable, while the bound $|\hat{E}| \leq c_3\, p\, s_{\max}$ keeps the search space manageable. Within this space, the orientation conditions—including a positive underfitting risk gap $\Delta_{\mathrm{miss}} > 0$, superset non-improvement, two-sided bounds on admissible population risks, and the direct-fit gain control in Assumption 6.6—ensure that an incorrect orientation raises the oracle direct-fit empirical score. Underfitted parent sets are detectable because the missing-parent risk gap is amplified in the tail, while spurious parents are excluded by the EBIC penalty. Together, these properties allow the orientation step to simultaneously discard false positives from the skeleton and recover the correct edge directions.

6.4.2 Consistency of score-based orientation

The population risk $R^*_{j|A}(u)$ in Section 6.2.2 optimises jointly over all offsets in the envelope class $\mathcal{H}_A$. To match that target, define the direct-fit empirical risk

$$\widehat{R}^{\mathrm{dir}}_{j|A}(u) := \inf_{\substack{c_0\in\mathbb{R}\\ c_{m\to j}\in\mathbb{R},\ m\in A}} \frac{1}{k}\sum_{t=1}^{k}\left|Z_{tj} - \max\left\{c_0, \max_{m\in A}\left(Z_{tm}+c_{m\to j}\right)\right\}\right|, \tag{6.10}$$

with the usual convention for $A=\emptyset$. Write $\widehat{f}^{\mathrm{dir}}_j(A) = \frac{k}{2}\log \widehat{R}^{\mathrm{dir}}_{j|A}(u)$ and

$$\widehat{S}^{\mathrm{dir}}(G) = \sum_{j=1}^{p}\left\{\widehat{f}^{\mathrm{dir}}_j\left(\mathrm{pa}_G(j)\right) + \frac{1}{2}\left(\log k + 2\gamma_{\mathrm{EBIC}}\log p\right)\left|\mathrm{pa}_G(j)\right|\right\}.$$

The results below concern the global minimiser of this direct-fit score. They do not identify the componentwise quantile offsets in Algorithm 6.1 with the joint optimum in (6.10).

Lemma 6.1 (Nodewise uniform empirical direct-fit risk convergence)**.** *Assume that, uniformly over all admissible pairs (j,A), the empirical tail prediction risk $\widehat{R}^{\mathrm{dir}}_{j|A}(u)$ satisfies*

$$Pr\left(\left|\widehat{R}^{\mathrm{dir}}_{j|A}(u) - R^*_{j|A}(u)\right| > \epsilon\right) \le c_1\exp\left\{-c_2\,k\min(\epsilon^2,\epsilon)\right\},$$

for some constants $c_1, c_2 > 0$ and all $\epsilon > 0$. Under the conditions of Theorem 6.3 and $\log p = o(k)$,

$$\max_{1\le j\le p}\sup_{\substack{A\subseteq \mathrm{nb}(j;\hat{E}_{\mathrm{skel}})\\ |A|\le s_{\max}}}\left|\widehat{R}^{\mathrm{dir}}_{j|A}(u) - R^*_{j|A}(u)\right| = O_p\left(\sqrt{\frac{(s_{\max}+1)\log p}{k}}\right).$$

The proof is deferred to Appendix C.5.

Assumption 6.6 (Equal-risk superset direct-fit gain control). Let $\widehat{f}^{\mathrm{dir}}_j(A) = \frac{k}{2}\log\widehat{R}^{\mathrm{dir}}_{j|A}(u)$. We assume the following uniform nested-model regularity condition: for all admissible pairs $A^*\subsetneq A$ in $\mathcal{A}_j^{c_3}$ satisfying $R^*_{j|A}(u) = R^*_{j|A^*}(u)$, the direct-fit gain from adding $B = A\backslash A^*$ obeys

$$\max_{1\le j\le p}\sup_{\substack{A^*\subsetneq A,\ A,A^*\in\mathcal{A}_j^{c_3}\\ R^*_{j|A}(u)=R^*_{j|A^*}(u)}}\frac{\widehat{f}^{\mathrm{dir}}_j(A^*) - \widehat{f}^{\mathrm{dir}}_j(A)}{|B|} = O_p(1).$$

Equivalently, $\widehat{f}^{\mathrm{dir}}_j(A^*) - \widehat{f}^{\mathrm{dir}}_j(A) = O_p(|B|)$ uniformly over equal-risk supersets.

In practice, the above means that adding $|B|$ spurious parents to the true parent set A^* can improve the direct empirical fit by at most $O_p(|B|)$, which the EBIC penalty, growing as $|B| \log p$, dominates in high dimensions. Intuitively, the fit gain from spurious parents is too small to offset the dimension-dependent penalty.

Theorem 6.4 (Consistency of the oracle direct-fit global empirical score minimizer)**.** *Assume the conditions of Theorem 6.3. For each node j, let $\mathcal{A}_j^{c_3}$ denote the class of admissible parent sets $A \subseteq V \setminus \{j\}$ with $|A| \le s_{\max}$ such that $A \subseteq \mathrm{nb}(j; E)$ for some $E \supseteq E^*_{\mathrm{skel}}$ satisfying $|E| \le c_3 p s_{\max}$. Suppose the following conditions hold at the chosen threshold u: (i) nodewise uniform underfitting gap,*

$$\inf_{1\le j\le p} \inf_{\substack{A\in\mathcal{A}_j^{c_3}\\ A\not\supseteq \mathrm{pa}^*(j)}} \left\{R^*_{j|A}(u) - R^*_{j|\mathrm{pa}^*(j)}(u)\right\} \ge \Delta_{\mathrm{miss}} > 0;$$

(ii) superset non-improvement, $R^*_{j|A}(u) = R^*_{j|\mathrm{pa}^*(j)}(u)$ for all $A \supseteq \mathrm{pa}^*(j)$ in $\mathcal{A}_j^{c_3}$ (verified for the recursive max-linear benchmark by Proposition C.1 in Appendix C.3); and (iii) two-sided admissible-risk bounds,

$$0 < r_{\min}(u) \le \inf_{1\le j\le p} \inf_{A\in\mathcal{A}_j^{c_3}} R^*_{j|A}(u) \le \sup_{1\le j\le p} \sup_{A\in\mathcal{A}_j^{c_3}} R^*_{j|A}(u) \le r_{\max}(u) < \infty.$$

Assume additionally Assumption 6.6 and

$$s_{\max}\sqrt{\frac{\log p}{k}} = o(1).$$

Let

$$\widehat{G}^{\mathrm{dir}} \in \arg\min_{G\in\mathcal{G}(\hat{E}_{\mathrm{skel}})} \widehat{S}^{\mathrm{dir}}(G)$$

be any global direct-fit EBIC score minimizer over DAGs compatible with $\hat{E}_{\mathrm{skel}}$. Then

$$Pr\big(\widehat{G}^{\mathrm{dir}} = G^*\big) \to 1 \qquad \text{as } k, p \to \infty.$$

The proof is deferred to Appendix C.7.

6.4.3 Robustness to latent confounding via proxy adjustment

We now consider the effect of residual confounding on the score-based orientation step. The skeleton stage already incorporates the proxy as an unpenalized regressor (Section 6.3.1), and Theorem 6.3 guarantees that true edges are retained under this adjustment. We therefore focus on the residual effect of imperfect proxy partialling-out on the orientation step, where the concern is that remaining confounding-induced dependencies could distort the population score ordering.

Assumption 6.7 (Vanishing residual confounding). Let P be the proxy introduced in Section 6.3.1. For each node j, let $\tilde{Z}_j = Z_j - \Pi_P(Z_j)$ denote the tail-domain residual after projecting onto the span of P (e.g., $\Pi_P(Z_j) = \arg\min_a \sum_t (Z_{t,j} - aP_t)^2$ on the tail sample). Let $\beta^\star_{ji}$ denote the corresponding population regression coefficient of $\tilde{Z}_j$ on $\tilde{Z}_i$ under the tail-limit working model.

Let $\mathcal{E}_{\text{conf}}$ denote pairs (i, j) that are d-separated in the true DAG $\mathcal{G}^\star$ but share the latent common driver U (i.e., purely confounded pairs). We assume that the residual confounding leakage satisfies

$$|\beta^\star_{ji}| \;\le\; \gamma_{\text{conf}}(u) \quad \text{for all } (i, j) \in \mathcal{E}_{\text{conf}},$$

where $\gamma_{\text{conf}}(u) \to 0$ as $u \to \infty$. For high-dimensional consistency, we further assume that $\gamma_{\text{conf}}(u)$ vanishes at a rate faster than the score-separation scale, for example $\gamma_{\text{conf}}(u) = o(k^{-1/2})$ or more generally $o(\Delta_{\text{miss}})$ when the nodewise population risks satisfy the uniform underfitting gap $\Delta_{\text{miss}} > 0$. This ensures that confounding-induced spurious correlations are asymptotically negligible compared to genuine causal signals.

Assumption 6.7 is intended to capture a low-dimensional common-shock mechanism rather than an arbitrary proxy choice. It is plausible when the latent driver enters many nodes through a shared monotone factor and the proxy is itself a tail-informative surrogate for that factor, so that partialling out P removes most of the common variation while leaving local propagation structure in the residuals (Pasche et al., 2023). The assumption does not require the proxy to identify the latent driver exactly; rather, it requires the proxy to absorb enough of the common tail component that residual pairwise dependence among purely confounded nodes becomes asymptotically negligible relative to the

orientation signal. Concrete proxy constructions satisfying this requirement are discussed in Section 6.6. Under Assumption 6.7, residual confounding does not overturn the population score ordering: the conditions of Theorem 6.4 remain satisfied, and the oracle direct-fit consistency guarantee continues to hold.

6.5 Simulation study

6.5.1 Simulation setup

We assess the finite-sample performance of the proposed framework under recursive max-linear structural equation models with heavy-tailed noise and latent confounding. The simulation study is organized around two questions, namely how performance changes with increasing dimension and how sensitive the procedure is to increasing confounding intensity. In each setting, results are summarized over 50 Monte Carlo replicates and evaluated through skeleton recovery, edge orientation, and overall DAG recovery.

For each replicate, we generate a directed acyclic graph $G = (V, E)$ from a Barabási–Albert construction (Barabási and Albert, 1999), with the attachment parameter and confounding design varied across experiments. We use this construction as a controlled sparse-graph stress test with heterogeneous neighbourhood sizes and hub nodes, allowing the method to be assessed on high-degree as well as low-degree parts of the graph. The model is generated by preferential attachment: nodes are added sequentially, and each new node forms m edges to existing nodes, favouring those that already have higher degree. The parameter m therefore controls the graph density. In the implementation, these edges are directed from the newly added node to the selected existing nodes, so the resulting graph is acyclic. Given G, the observed variables are generated from the recursive max-linear model

$$X_j = \max\left(\bigvee_{i \in \mathrm{pa}(j)} \{c_{ij} X_i\}, \bigvee_{l \in \mathrm{conf}(j)} \{b_{lj} U_l\}, \epsilon_j \right), \qquad j = 1, \ldots, p, \tag{6.11}$$

where $\mathrm{pa}(j)$ is the parent set of node j, $\mathrm{conf}(j)$ indexes latent confounders acting on node j, and the innovations ϵ_j and latent drivers U_l are independent standard Fréchet variables. The edge coefficients c_{ij} are sampled from $[0.6, 0.9]$, and the confounding coefficients b_{lj} from $[0.5, 0.8]$. For each latent driver U_l, we additionally generate a proxy $P_l = \max(a_l U_l, \eta_l)$, with $a_l \sim U[0.6, 0.9]$ and independent Fréchet noise η_l. Here q denotes the number of latent confounders, and conf_s denotes the number of observed nodes affected by each confounder.

After simulating n observations, each margin is transformed to unit Pareto scale by the empirical rank transform. We then follow the multivariate GPD exceedance rule and retain observations satisfying $\max_{1\le j\le p} Y_{tj} > u$, where u is defined by the common quantile level $q_{\mathrm{conf}} = 1 - n^{0.7}/n$. Equivalently, we select rows in which at least one margin exceeds its marginal q_{conf} threshold. Let k_{exc} denote the realized number of selected exceedances under this rule; k_{exc} is data-adaptive and not fixed a priori. We then work with the log-tail coordinates $Z = \log Y - \log u$ as in Section 6.3.1. Competing methods and experiment-specific parameter settings are introduced in the corresponding subsections.

6.5.2 High-dimensional scaling

We study scaling with fixed sample size $n = 1000$ and dimension $p \in \{20, 50, 100, 200\}$. Graphs are generated from Barabási–Albert models with $m \in \{1, 2\}$ and the default generator setting $q = \lfloor 0.1p \rfloor$, and confounding exposure width is set to $\mathrm{conf}_s = 0.2p$ (that is, $\{4, 10, 20, 40\}$ across dimensions). The tail sample is selected by the exceedance rule above, so the realized k_{exc} is data-adaptive. We use $\lambda = C\sqrt{\log p / k_{\mathrm{exc}}}$ with pre-specified $C = (0.5, 1.0, 1.5, 2.0)$ for $p = (20, 50, 100, 200)$, respectively, and $\gamma_{\mathrm{EBIC}} = (10, 10, 10, 20)$, with both choices fixed a priori and a larger penalty at $p = 200$ to enforce stronger complexity control in the largest search space.

Four methods are compared. The S3ME skeleton is the proxy-adjusted nodewise ℓ_1 estimator of Section 6.3.1, evaluated on its own as a screening device. S3ME denotes the full two-stage procedure comprising greedy EBIC-score orientation and pruning on the S3ME skeleton. EASE (Gnecco et al., 2021b) is a tail-conditional independence test applied to orient the same recovered skeleton, so that the comparison isolates the contribution of the orientation score; it uses $k = \lfloor n^{0.7} \rfloor$ tail observations and the same marginal trans-

formation, with no pruning step. FGES (Chickering, 2002) provides a standard body-data baseline using the Gaussian BIC score on the full sample of $n = 1000$ observations; it is run independently with penalty discount 1.0 and faithfulness assumed. Figure 6.1 summarizes F_1 and SHD, while Precision and Recall are reported in the Appendix. For EASE, FGES, and the full S3ME procedure, the directed F_1 and SHD are evaluated on the true skeleton.

All methods degrade as p grows, but the rate of degradation differs markedly. FGES shows a pronounced deterioration in overall recovery quality, with lower F_1 and larger SHD as dimension increases, because its Gaussian BIC score cannot distinguish tail propagation from body-level correlation in heavy-tailed data. The S3ME skeleton remains a liberal screening device, so its F_1 stays modest (mean skeleton F_1 of $0.27, 0.28, 0.25, 0.21$ for $p = 20, 50, 100, 200$) and its SHD increases with dimension as confounding-induced false positives accumulate. At $p = 200$, the proposed orientation step still outperforms EASE, with mean DAG F_1 of 0.66 for S3ME versus 0.50 for EASE, and a correspondingly wider SHD gap in favour of S3ME.

Confounding increasingly dominates skeleton errors in high dimensions, with the confounding share among skeleton false positives, $\text{ConfFP}_{\text{frac}}$, increasing from 0.08 at $p = 20$ to 0.92 at $p = 200$. Because EASE orients but does not prune the shared skeleton, its DAG-stage ConfFP tracks the skeleton-level values exactly. Across $p = 20, 50, 100, 200$, DAG-stage ConfFP is $(0.0, 1.2, 6.1, 16.1)$ for S3ME, compared with $(7.8, 92.0, 312.5, 707.2)$ for EASE and $(4.1, 47.9, 245.8, 1086.9)$ for FGES. At $p = 200$, this corresponds to roughly 44× and 67× reductions relative to EASE and FGES, respectively. FGES accumulates even more confounding FPs because it lacks both a tail selection mechanism and a confounding-aware screening step. This quantitative gap is consistent with the design of the orientation score: the EBIC penalty discourages spurious parents whose empirical fit improvement does not justify the complexity cost, while the tail asymmetry criterion favours directions with lower tail prediction error.

6.5.3 Sensitivity to confounding intensity

This experiment is a controlled stress test that uses the latent-driver signal as proxy input (oracle-style alignment). We fix $p = 100$, set $m = 1$, use one latent confounder ($q = 1$), and vary confounding exposure width via $\text{conf}_s \in \{0, 0.2p, 0.5p, p\}$. All other settings are fixed.

Table 6.1 compares three methods at the DAG level, namely the full proposed procedure (S3ME), an ablation that removes the proxy adjustment from both stages (No proxy), and EASE applied to the S3ME skeleton. For symmetry clarification, S3ME and EASE share the proxy-adjusted skeleton, while No proxy is an end-to-end ablation using its own no-proxy skeleton. The table reports directed-stage F_1 and confounding-induced false positives (ConfFP).

Table 6.1: DAG-level confounding sensitivity for $p = 100$, $m = 1$, and 50 Monte Carlo replicates, comparing S3ME, its no-proxy ablation, and EASE on the S3ME skeleton. The confounding exposure width conf_s is the number of observed nodes affected by the latent driver. $F_1 = 2PR/(P + R)$, with $P = \text{TP}/(\text{TP} + \text{FP})$ and $R = \text{TP}/(\text{TP} + \text{FN})$. ConfFP counts estimated edges absent from the true DAG whose endpoints share a latent confounder.

	S3ME		No proxy		EASE	
conf_s	F_1	ConfFP	F_1	ConfFP	F_1	ConfFP
0	0.831	0.0	0.847	0.0	0.671	0.0
0.2p	0.836	0.1	0.818	6.4	0.467	74.7
0.5p	0.837	0.5	0.767	18.9	0.294	206.9
p	0.837	1.3	0.701	39.1	0.178	402.2

The pattern is sharp. Without confounding ($\text{conf}_s = 0$), the no-proxy ablation is slightly better ($F_1 = 0.847$ vs 0.831). The crossover appears immediately at $\text{conf}_s = 0.2p$, where S3ME overtakes no-proxy (0.836 vs 0.818), and the advantage then widens monotonically with confounding strength. At $\text{conf}_s = p$, S3ME attains $F_1 = 0.837$ versus 0.701 for no-proxy and 0.178 for EASE. The same gradient appears in ConfFP, with values 1.3 (S3ME), 39.1 (no-proxy), and 402.2 (EASE) at $\text{conf}_s = p$. This confirms that the proxy term is not uniformly beneficial, but is critical for robustness once confounding is present. The ablation further indicates that the main gain is a cleaner candidate structure, as S3ME keeps far fewer confounding-induced edges in the final DAG at every confounded setting

(0.1 vs 6.4, 0.5 vs 18.9, and 1.3 vs 39.1), which aligns with its higher F_1. We emphasize that this phase uses latent-driver proxy input as a controlled stress test, so the result should be read as robustness evidence under aligned proxy information rather than as a full observational-proxy guarantee.

Additional robustness checks are reported in Appendices C.8 and C.9. Appendix C.9 compares performance across Erdős–Rényi and Barabási–Albert graph topologies at matched edge density, showing that the main conclusions hold for ER graphs while the lower recall observed under BA with $m = 2$ is attributable to hub-induced neighbourhood complexity rather than a failure of the method. Appendix C.8 examines sensitivity to the exceedance threshold β, confirming that DAG-level performance is stable across $\beta \in \{0.5, \ldots, 0.9\}$ and that the default $\beta = 0.7$ lies in a plateau of near-optimal performance.

6.6 Applications

We illustrate the method on two real-data settings with different validation regimes, namely a river network with a physically interpretable flow structure and a financial network without known ground truth.

6.6.1 Danube river network

We first apply the proposed framework to extreme river discharge data from the upper Danube basin. We analyze daily summer (June–August) discharge measurements from $n = 428$ observations at 31 gauging stations provided by the Bavarian Environmental Agency, restricted to the common observation period 1960–2010 (Asadi et al., 2015). After standard temporal declustering and tail preprocessing (Appendix C.10), the threshold $q_{\text{conf}} = 0.90$ yields $k = 117$ tail observations. This setting is useful because the river topology, derived from the known catchment structure of the upper Danube (Asadi et al., 2015; Engelke and Hitz, 2020), provides a physically interpretable benchmark for the direction of extremal propagation.

A central difficulty in this application is the presence of regional meteorological forcing: large-scale storm systems can drive extreme discharges at multiple stations simultaneously, inducing strong symmetric tail dependence that obscures the directional propagation along the river network. To attenuate this common component, we construct a proxy

series $P_t = \sum_{j=1}^{p} \mathbb{I}(Z_{t,j} > 0)$, which records the spatial extent of simultaneous exceedances across the network and serves as a scalar summary of region-wide meteorological activity. We then apply the two-stage procedure to the preprocessed data, using $\lambda = \sqrt{\log(p)/k}$ for skeleton estimation and $\gamma_{\text{EBIC}} = 10$ in the orientation step.

Figure 6.2 compares the estimated extremal DAG with the known physical river-flow graph. Against the 30 true edges in the upstream-to-downstream topology, the method recovers 24 skeleton edges with precision 0.88 and recall 0.70 ($F_1 = 0.78$). Among the 21 correctly identified edges, 17 (0.81) receive the correct direction and 4 (0.19) are reversed. The reversed edges occur at tributary confluence points, where the max-linear signal from the main stem can dominate the tributary contribution and make the direction ambiguous from tail behaviour alone. The 9 missing edges are predominantly short-range connections along the main stem and at smaller tributary junctions, where the signal from upstream propagation is weaker relative to the noise from local catchment variability. The 3 false positive edges, by contrast, connect geographically distant stations and likely reflect residual co-exceedance not fully absorbed by the proxy. The proxy adjustment suppresses many dense associations that would otherwise be induced by region-wide storm events, while retaining the dominant propagation pathways.

6.6.2 S&P 500 tail risk network

We next apply the method to daily equity returns from S&P 500 constituents. The data are obtained from Stooq[1] over the period January 2005 to March 2025, yielding approximately $n = 5{,}093$ trading days and $p = 103$ stocks spanning all eleven GICS sectors. Loss variables are defined by $L_{t,j} = -r_{t,j}$, where $r_{t,j}$ denotes the log-return of stock j on day t.

A main difficulty in this application is the presence of strong market-wide tail dependence during periods of systemic stress. To attenuate this common component, we construct a three-dimensional proxy $\boldsymbol{S}_t = (\text{VIX}_t, -r_t^{\text{SP500}}, \Delta\text{DGS10}_t)^\top$, where VIX_t captures implied market volatility (a forward-looking measure of expected tail risk), $-r_t^{\text{SP500}}$ captures broad equity drawdowns (the contemporaneous market-wide loss), and ΔDGS10_t captures daily changes in the 10-year Treasury yield (a proxy for flight-to-quality shifts that can induce simultaneous tail dependence across equity sectors). We retain $k = 4{,}120$

1. https://stooq.com

tail observations using the adaptive threshold $q_{\text{conf}} = 1 - n^{0.7}/n$; see Appendix C.10 for data cleaning, preprocessing, and tuning details. The skeleton is then estimated by proxy-adjusted nodewise extremal LASSO with $\lambda = c\sqrt{\log(p+1)/k}$, using $c = 5$, and edges are oriented with $\gamma_{\text{EBIC}} = 10$ and maximum in-degree 10. The resulting graph contains 113 directed edges among 103 nodes.

Figure 6.3 shows the estimated network. The graph exhibits clear within-sector structure together with a smaller number of cross-sector transmission pathways. Utilities form the most prominent cluster, with AEP (out-degree 5) and XEL (out-degree 5) appearing as major transmitters and several downstream links concentrated within the same sector. Real-estate names occupy a more intermediate position, connecting the Utilities cluster to parts of the Financial and Consumer sectors. The estimated graph also identifies a technology subnetwork involving KLAC, AMAT, LRCX, and NVDA, which is consistent with concentrated dependence within semiconductor-related supply chains. More broadly, the hub structure, with a small number of high out-degree nodes concentrated in Utilities, Financials, and Materials, suggests that tail risk propagation during market stress is channelled through a few sector-specific transmitters rather than diffusing uniformly across the market. These patterns should not be interpreted as definitive ground-truth causal claims in a market without known validation structure. Rather, the main value of the application is that the estimated graph is sparse, sectorally interpretable, and qualitatively aligned with plausible channels of tail risk propagation under common market stress.

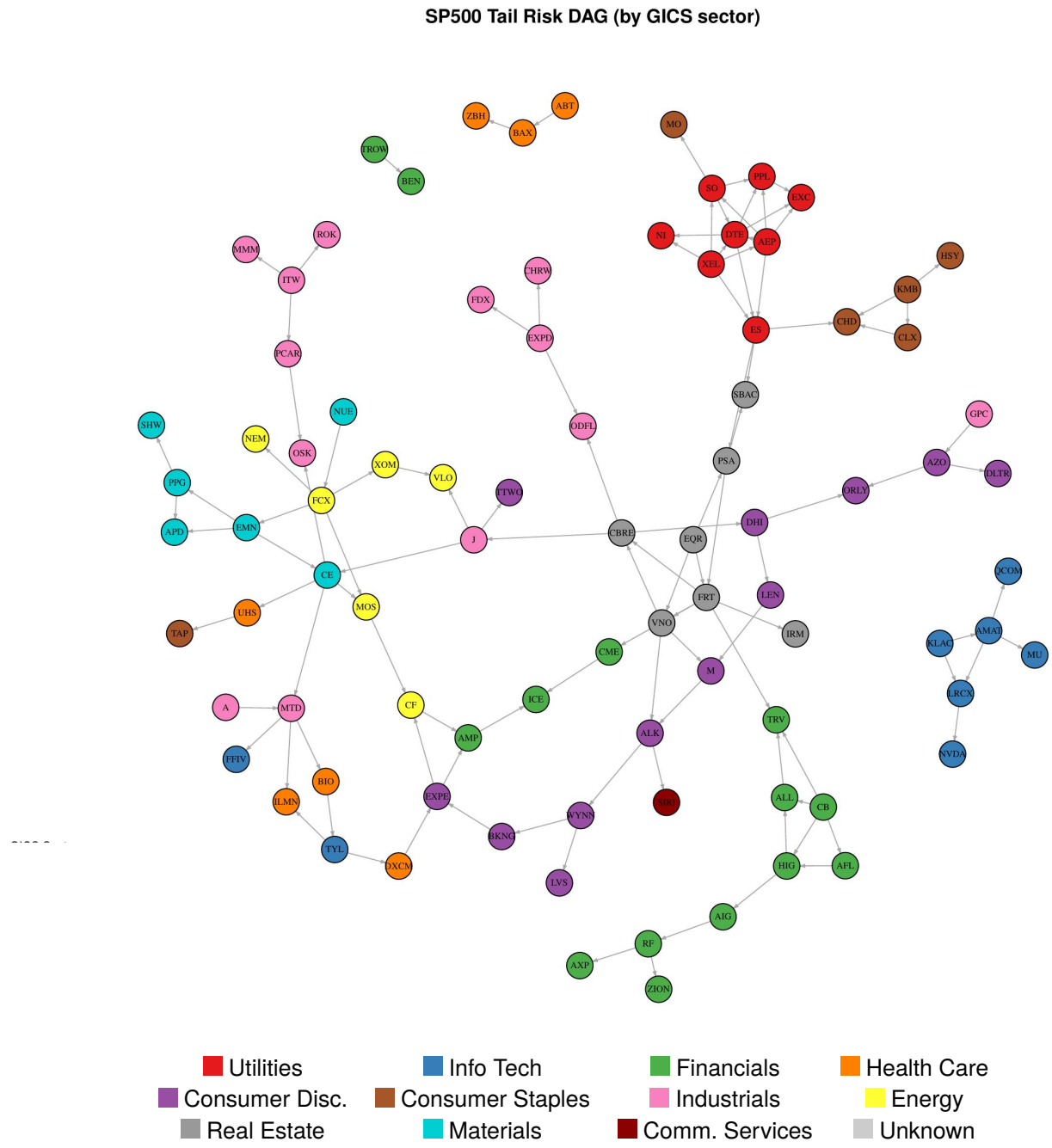

Figure 6.3: Estimated tail risk network among S&P 500 constituents. Nodes are colored by GICS sector, and directed edges indicate inferred directions of tail risk propagation.

6.7 Conclusion

We have proposed a two-stage procedure for causal discovery in multivariate extremes, based on tail-induced asymmetry. The first stage screens for a sparse candidate skeleton with proxy-adjusted extremal nodewise regression. The second stage orients candidate edges by score minimisation under a max-linear envelope with ℓ_1 loss. The objective is not to recover bulk dependence, but to extract directional information from the tail regime.

The evidence is consistent across theory, simulation, and data analysis. At population level, Theorem 6.1 establishes forward–backward separation of tail prediction error in a canonical bivariate max-linear setting. In finite samples, the simulation results in Section 6.5 show that proxy adjustment reduces false positives under latent common shocks while retaining useful directional recovery. In the Danube application, where a physical flow topology is available, 17 of 21 correctly identified skeleton edges are oriented correctly (81%). In the S&P 500 analysis, where no ground-truth DAG is available, the estimated network is interpreted as a sparse and sectorally coherent representation of tail-risk transmission, rather than as validated causal truth.

The current guarantees are deliberately scoped. The identifiability argument is developed for a bivariate max-linear model. The high-dimensional results rely on a working Hüsler–Reiss approximation and a population score-separation condition. The superset non-improvement property is verified for the max-linear subclass (Proposition C.1, Appendix C.3), not yet for the full class in Assumption 6.1. Moreover, Theorem 6.4 concerns the oracle global minimiser of the direct-fit envelope score, whereas the implemented orientation step uses componentwise quantile-calibrated offsets and greedy search. The theorem does not cover either computational approximation. Proxy adjustment should therefore be viewed as a structural safeguard against common-shock dependence, not as a full correction for hidden confounding.

In terms of future work, directions appear most immediate. First, uncertainty quantification for learned edges (for example, bootstrap stability or tail-focused subsampling) would improve practical interpretability. Second, computational refinements are needed when the candidate skeleton is dense and orientation search becomes costly. Third, extending orientation guarantees beyond the max-linear envelope, together with richer domain-specific proxy design, would broaden applicability.

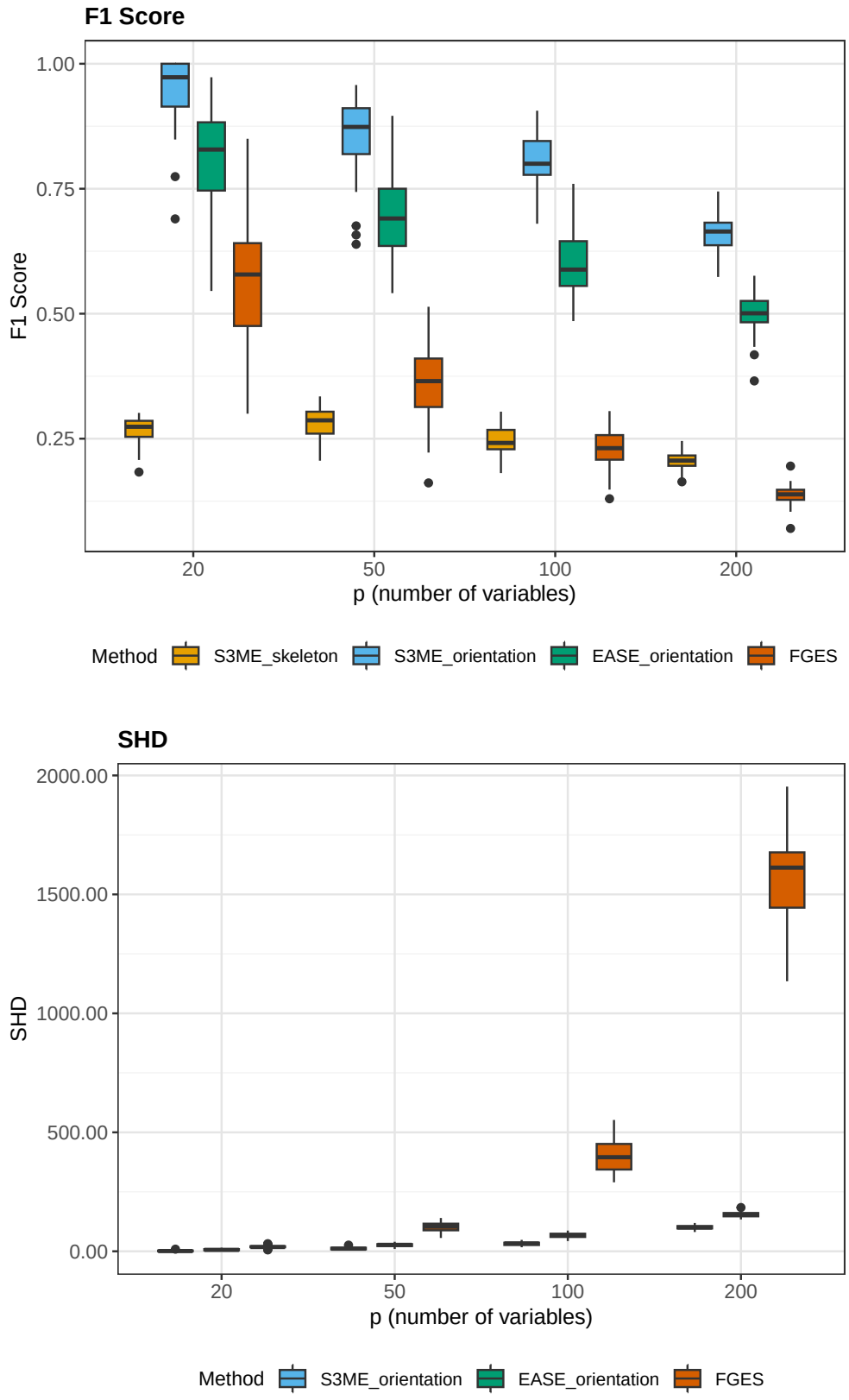

Figure 6.1: High-dimensional scaling results for $m = 1$, $n = 1000$, and 50 Monte Carlo replicates. The upper and lower panels show F_1 and structural Hamming distance (SHD), respectively, across $p \in \{20, 50, 100, 200\}$ for the S3ME skeleton, EASE, FGES, and the full S3ME procedure. For EASE, FGES, and the full S3ME procedure, the directed F_1 and SHD are evaluated on the true skeleton. For precision $P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and recall $R = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, $F_1 = 2PR/(P+R)$. The implemented SHD is the number of directed adjacency disagreements, $\mathrm{FP} + \mathrm{FN}$.

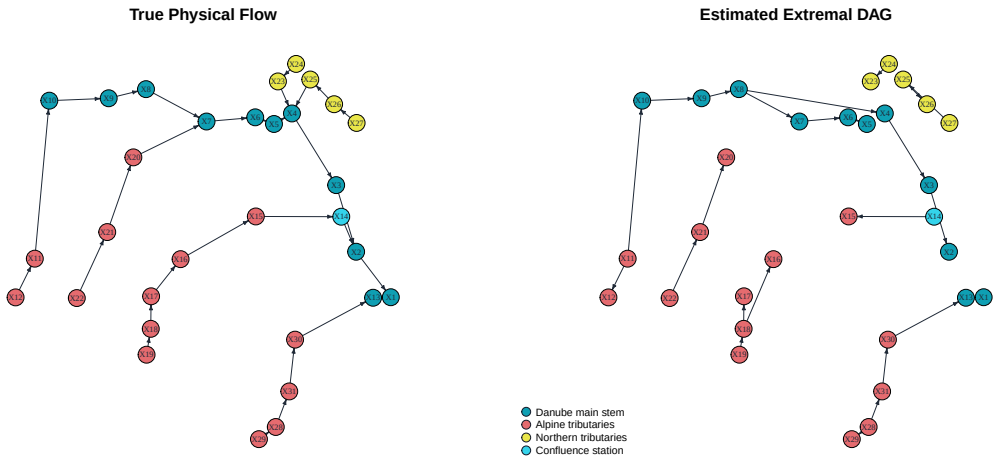

Figure 6.2: Upper Danube comparison. Left Panel shows the physical river-flow graph, while the right panel shows the estimated extremal DAG, both on the same geographic layout. Node colors indicate basin groups: Danube main stem (teal), Alpine tributaries (salmon), northern tributaries (yellow), and the confluence station (cyan).

Chapter 7
Conclusion

7.1 Contributions of the thesis

This thesis makes contributions through three distinct but related tasks: attributing rare events to their causes, estimating treatment effects on extreme outcomes, and discovering the structural relationships that drive extremes. Each task was treated in a dedicated chapter, summarised below.

Chapter 4 investigated the sensitivity of extreme event attribution conclusions to statistical tail specification. The chapter compared the attribution ratio and the probability of necessary causation across three parametric tail models spanning asymptotic dependence, asymptotic independence, and hybrid regimes. Its HW-based simulations showed material differences under the marginal and dependence perturbations considered, while the two precipitation applications produced model-dependent attribution maps. The contribution is a sensitivity framework showing how tail-model choice affects statistical attribution and motivating joint assessment of marginal and dependence assumptions.

Chapter 5 developed the Tail-Calibrated Inverse Estimating Equation (TIEE) framework for estimating quantile treatment effects at extreme levels. The framework replaces the unstable pointwise inversion of standard quantile estimators with a globally solved estimating equation that integrates inverse probability weighting with generalised Pareto tail calibration. Identification was established under unconfoundedness and a tail overlap condition strictly weaker than standard support assumptions. Consistency and asymptotic normality were proved, with convergence rates that improve upon standard inverse probability weighted quantile estimators by a logarithmic factor at extreme levels. Simulation studies across heavy- and light-tailed regimes confirmed substantial bias and variance reductions, and an application to Austrian precipitation data illustrated the framework's capacity for causal attribution from observational records.

Chapter 6 addressed causal discovery in multivariate extremes, where standard graphical model methods are confounded by strong tail co-exceedance and latent common shocks. The chapter identified tail-induced asymmetry (the directional imbalance in extremal prediction risk) as a source of causal directionality, and built a two-stage framework around it: proxy-adjusted penalised regression for skeleton recovery, followed by score-based edge orientation via max-linear envelope prediction. A sure screening guarantee was established for skeleton recovery, and asymptotic identifiability was proved for edge orientation. Applications to river network flow and financial tail risk data demonstrated recovery of interpretable propagation structures in settings where bulk-based methods fail.

Taken together, the three chapters contribute a sensitivity-based critique of dependence specification in attribution, a tail-calibrated estimating-equation framework for treatment-effect estimation at extreme quantiles, and a two-stage procedure built on tail-induced asymmetry for causal discovery in multivariate extremes. The next section draws out the common role of tail assumptions across these three contributions.

7.2 Causality beyond the bulk

The core chapters address three distinct inferential tasks: attribution, estimation, and discovery. This section draws out a common thesis-level perspective across them. Tail behaviour enters each task differently: through marginal and dependence specification in attribution, through extrapolation in extreme treatment-effect estimation, and through extremal propagation in causal discovery.

In Chapter 4, the attribution ratio and probability of necessary causation depend on factual and counterfactual probabilities of multivariate extremes. The chapter compares three fitted tail models and shows how alternative marginal and dependence specifications produce different attribution estimates in the simulations and precipitation applications considered.

In Chapter 5, the causal target is an extreme quantile treatment effect. Estimation beyond the well-supported empirical range requires extrapolation, so GPD tail calibration enters the estimating equation together with the causal weighting structure. Extreme value theory thereby supports estimation of a causal quantity in a sparsely observed region of the outcome distribution.

In Chapter 6, the objective is structural rather than effect-based. The method uses extremal prediction risk and tail-induced asymmetry to learn causal direction from multivariate extremes. Tail observations supply the information used for skeleton recovery and edge orientation.

Each chapter retains its own estimand, assumptions, methodology, and theoretical analysis. Together, they support a common practical principle: tail assumptions should be assessed in relation to the causal target. Chapter 4 implements this through model comparison and diagnostics, Chapter 5 through tail-calibrated estimating equations, and Chapter 6 through causal structure learning in the tail domain. These are distinct methodological responses to a shared challenge posed by causal questions about extremes.

7.3 Open problems and future directions

The contributions of this thesis also come with clear limitations, and these limitations point directly to the next methodological steps. Taken together, they sketch where a broader programme of causal inference for extremes must continue to develop.

Attribution. Chapter 4 focused on three parametric dependence models: the multivariate generalised Pareto distribution, the exponential factor copula model, and the Huser–Wadsworth model, chosen to span the range from asymptotic dependence to asymptotic independence. While these models cover the dependence regimes most commonly encountered in environmental applications, they do not exhaust the space of possible extremal dependence structures, and in particular they cannot represent dependence whose spatial and temporal structure shifts under changing climate conditions. A natural next

step is therefore to move beyond fixed parametric families and adopt neural network parameterisations of extremal dependence learned directly from data. Whether attribution conclusions remain stable under such richer dependence models is itself an open empirical question.

Estimation. The TIEE framework leaves several natural extensions open. It can be extended from binary to continuous treatments via generalised propensity scores, to spatial settings in which the estimand is a random field rather than a scalar, and to a double machine learning formulation in which Neyman orthogonalisation neutralises the slow convergence rates of flexible nuisance estimators. The last direction is the most consequential, since the slower rates of data-driven nuisance estimation compound with the additional variability introduced by extrapolating into the tail, and a principled treatment of this interaction is a precondition for scaling the framework to high-dimensional observational settings.

Discovery. Chapter 6 established identifiability and consistency results for a recursive max-linear structural equation model. This gives a clean theory for tail-induced asymmetry, but it does not yet cover broader recursive extremal models with additive or nonlinear structure. The skeleton recovery stage also relies on a Hüsler–Reiss working model, and the proxy adjustment for latent confounding attenuates symmetric co-exceedance rather than removing it completely. Extending the theory beyond the max-linear setting, and to dynamic or time-series extremes, is an important direction for future work.

Unified inference. An overarching limitation of the thesis is that the three strands have not yet been merged into a single inferential pipeline. Chapter 4 studies dependence sensitivity in attribution, Chapter 5 integrates causal adjustment with tail extrapolation, and Chapter 6 recovers causal structure from extremal data, but the output of one chapter is not yet fed directly into the next. A longer-term goal is therefore a unified procedure that discovers structure from tail data, estimates extreme causal effects conditional on that structure, and propagates uncertainty through to attribution statements. Achieving that integration would turn the conceptual arc of this thesis into a single operational framework.

7.4 Closing perspective

The central conclusion of this thesis is that causal inference in the tails cannot always be treated as a simple extension of causal inference in the centre of a distribution. Across the problems studied here, tail geometry introduces extrapolation, dependence, and identifiability issues that are not captured by methods aimed primarily at typical outcomes. The resulting conclusions can be sensitive to tail-modelling assumptions, and that sensitivity should be assessed explicitly rather than assumed to be negligible.

The methods developed here for attribution under tail-model uncertainty, for extreme quantile treatment effects, and for causal discovery from tail asymmetry are all motivated by that observation. In each case, extreme value modelling and causal reasoning are developed jointly because the relevant tail assumptions enter the estimand, estimator, or structural model. This is a methodological lesson from the settings analysed in the thesis, not a claim that every causal problem involving an extreme outcome has the same structure.

Many problems remain open. The broader lesson supported by these chapters is that, when a causal target lies in the tail, tail assumptions should enter the analysis from the outset and their influence on the conclusions should be reported explicitly.

Appendices

A Appendix material for Chapter 4

A.1 Profile likelihood for marginal HW model

Figure A.1 displays the profile of the marginal negative log-likelihood for δ at a representative location in the European precipitation application. The flat shape near the upper boundary illustrates the limited information available for estimating this parameter.

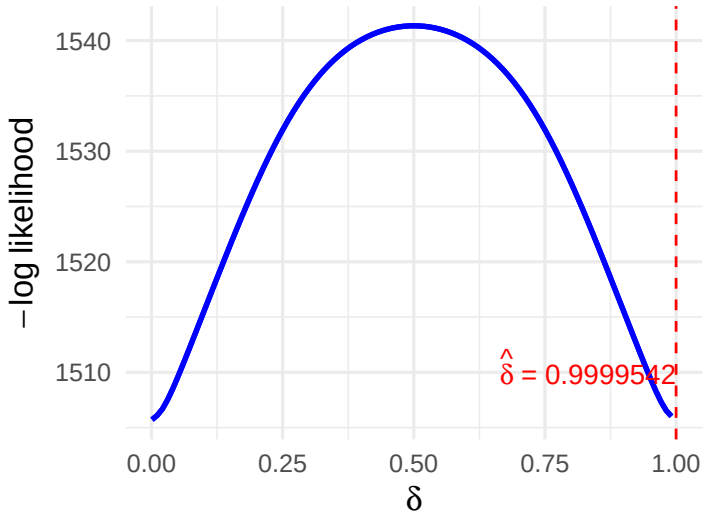

Figure A.1: Profile negative log-likelihood for the HW dependence parameter $\delta \in [0, 1]$ in (4.7), using the marginal fit at location 217 in the European precipitation application. The minimizer lies near the upper boundary; the comparatively flat profile there indicates weak information about δ at this location.

A.2 Tail dependence coefficients for the candidate dependence models

Here, we provide derivations for the conditional probabilities $\chi(u)$ for the eFCM and the HW models introduced in Section 4.2.

A.2.1 Derivation of $\chi(u)$ for the eFCM

We consider the exponential factor copula model (eFCM), where the dependence structure arises from a Gaussian copula with margins given by:

$$W_j = Z_j + \lambda_j^{-1} E, \quad j = 1, 2,$$

where $Z = (Z_1, Z_2)^\top \sim \mathcal{N}(0, \Sigma)$, with Σ having unit variances and correlation ρ, and $E \sim \text{Exp}(1)$, independent of Z. Let $F_j(w; \lambda_j)$ denote the marginal distribution of W_j, and define

$$z(u) = F_j^{-1}(u; \lambda_j),$$

so that $W_j \sim F_j$ and $F_j(z(u); \lambda_j) = u$. It can be shown that $\chi(u)$ can be expressed as

$$\chi(u) = \frac{1 - 2u + C_2(u, u)}{1 - u},$$

where C_2 is the bivariate copula. Using the eFCM, we have $\chi_{\text{eFCM}}(u) = \frac{1-2u+C_2^W(u,u)}{1-u}$, where C_2^W is the eFCM bivariate copula. From the supplementary material of Castro-Camilo and Huser (2020), $C_2^W(u, u)$ has the form:

$$C_2^W(u, u) = \Phi_2(z_1(u), z_2(u); \rho) - \sum_{j=1}^{2} \exp\left(\frac{\lambda_j^2}{2} - \lambda_j z_j(u)\right) \Phi_2(q_j, 0; \Omega_j),$$

where $\Phi_2(\cdot,\cdot;\rho)$ is the bivariate standard normal CDF with correlation ρ, q_j and Ω_j are functions of λ_1, λ_2 and ρ, provided in page 2 of the supplementary material of Castro-Camilo and Huser (2020), and $z_j(u) = F_j^{-1}(u;\lambda_j)$. Thus, the expression for $\chi_{\text{eFCM}}(u)$ becomes:

$$\chi_{\text{eFCM}}(u) = \frac{1 - 2u + \Phi_2(z_1(u), z_2(u); \rho) - \sum_{j=1}^{2} \exp\left(\frac{\lambda_j^2}{2} - \lambda_j z_j(u)\right) \Phi_2\left(q_j, 0; \Omega_j\right)}{1-u}$$

In the symmetric case ($\lambda_1 = \lambda_2 = \lambda$), the formula simplifies to:

$$\chi_{\text{eFCM}}(u) = \frac{1 - 2u + \Phi_2(z(u), z(u); \rho) - 2\exp\left(\frac{\lambda^2}{2} - \lambda z(u)\right) \Phi_2\left(q, 0; \Omega\right)}{1-u}$$

where:

$$q = \lambda(1-\rho), \quad \Omega = \begin{pmatrix} 2(1-\rho) & \sqrt{2(1-\rho)} \\ \sqrt{2(1-\rho)} & 1 \end{pmatrix}.$$

A.2.2 Derivation of $\chi(u)$ for the HW model

Recall that in the Huser and Wadsworth (HW) model, the process is defined as $X(s) = R^{\delta} W(s)^{1-\delta}$, $\delta \in (0,1)$, where $R \sim$ Pareto(1) and $W(s)$ is an asymptotically independent process with standard Pareto margins. Taking logarithms gives the additive form

$$\tilde{X}(s) := \log X(s) = \delta \tilde{R} + (1-\delta)\tilde{W}(s),$$

where $\tilde{R} = \log R \sim \text{Exp}(1)$, and $\tilde{W}(s) = \log W(s)$, with $\tilde{W}_j \sim \text{Exp}(1)$ and copula structure determined by the dependence in W.

Let F_X denote the marginal distribution of $X(s)$. Then the bivariate copula $C_2(u,u)$ associated with the HW model is given by

$$C_2(u,u) = \Pr\left(X_1 \le w(u), X_2 \le w(u)\right),$$

where $w(u) = F_X^{-1}(u)$.

From the log-scale representation, this can be written as

$$C_2(u,u) = \Pr\left(\tilde{X}_1 \leq \log w(u), \tilde{X}_2 \leq \log w(u)\right)$$
$$= \int_0^{\log w(u)/\delta} F_{\tilde{W}_1,\tilde{W}_2}\left(\frac{\log w(u) - \delta r}{1-\delta}, \frac{\log w(u) - \delta r}{1-\delta}\right) e^{-r}\,\mathrm{d}r,$$

where $F_{\tilde{W}_1,\tilde{W}_2}$ is the joint distribution of $(\tilde{W}_1, \tilde{W}_2)$.

Substituting into the standard formula for the conditional tail dependence function:

$$\chi(u) = \frac{1 - 2u + C_2(u,u)}{1-u},$$

we obtain:

$$\chi_{\mathrm{HW}}(u) = \frac{1 - 2u + \int_0^{\log w(u)/\delta} F_{\tilde{W}_1,\tilde{W}_2}\left(\frac{\log w(u) - \delta r}{1-\delta}, \frac{\log w(u) - \delta r}{1-\delta}\right) e^{-r}\,\mathrm{d}r}{1-u}$$

where $w(u) = F_X^{-1}(u)$ is the marginal quantile function of the HW model, $F_{\tilde{W}_1,\tilde{W}_2}$ is the bivariate distribution function of the log-transformed generator process, and δ controls the degree of asymptotic dependence.

A.3 Gaussian and IEVL copula comparison

Figure A.2 compares the density contours of the Gaussian and inverted extreme value logistic (IEVL) copulae in the upper tail region, as used in simulation sub-scenario 1.2.

A.4 Additional attribution maps for the CNRM application

Figure A.3 presents maps of attribution ratio point estimates (excluding confidence intervals) for the first application described in Section 4.4.1. By removing the visual influence of uncertainty, this simplified representation highlights the underlying spatial patterns, facilitating clearer interpretation and more direct comparison across models.

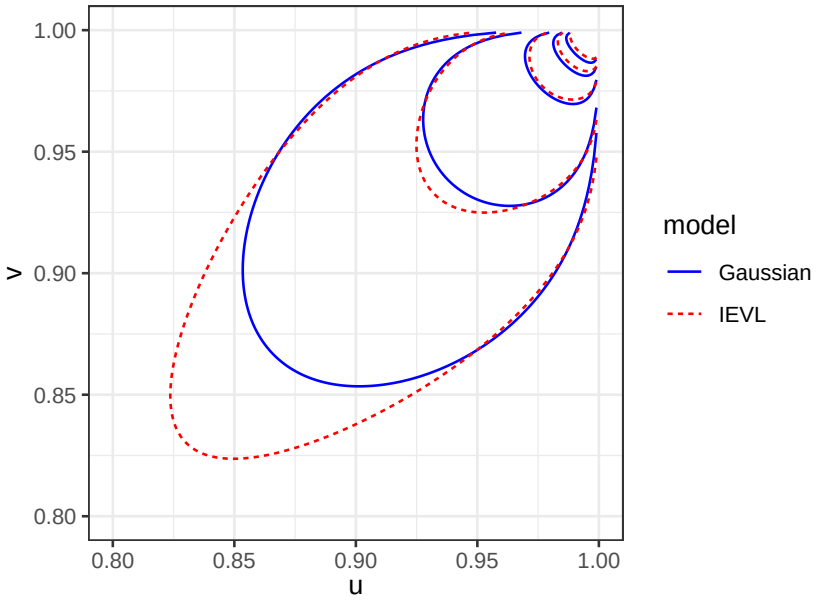

Figure A.2: Copula-density contours over $(u, v) \in [0.80, 0.999]^2$ for the Gaussian dependence model used by the HW data-generating process and the inverted extreme-value logistic (IEVL) model used as the misspecified dependence in Simulation 1.2 (Section 4.3.1.2). Higher contour density near $(1, 1)$ represents stronger simultaneous upper-tail concentration.

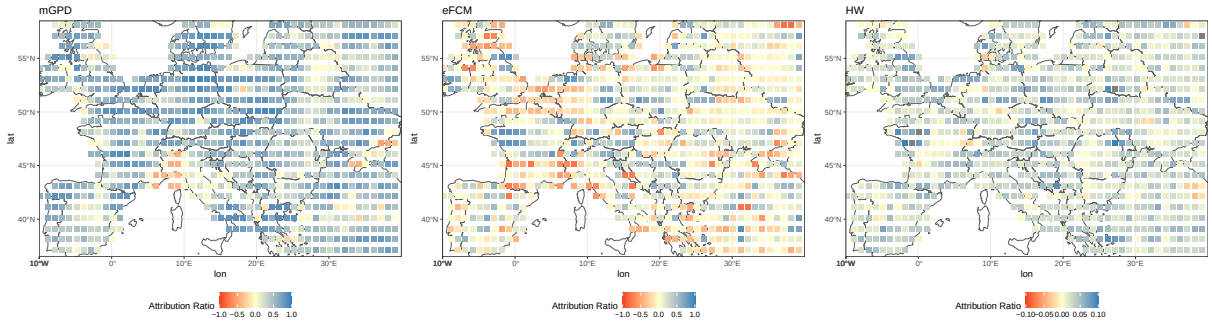

Figure A.3: Point estimates of the attribution ratio $\mathrm{AR} = (p_0 - p_1)/p_1$, with $p_t = \Pr\{\boldsymbol{w}^\top \boldsymbol{X}^{(t)} > v\}$ from (4.1), for the European precipitation application. Panels compare the mGPD, eFCM, and HW fits at the 5-year and 50-year return levels; confidence-interval opacity is omitted to isolate the fitted spatial point patterns.

A.5 Estimated PN values for the two data applications

As a complement to the AR maps presented in Sections 4.1 and 4.2 of Li et al. (2025), we also report estimates of the probability of necessary causation (PN). Figure A.4 displays the PN estimates for the European precipitation application using CNRM outputs, while Figure A.5 shows the corresponding results for the contiguous U.S. application based on ACCESS-CM2 outputs. Both figures present results for the 5-year and 50-year return levels across the three models. Overall, the conclusions align closely with those drawn from the AR maps.

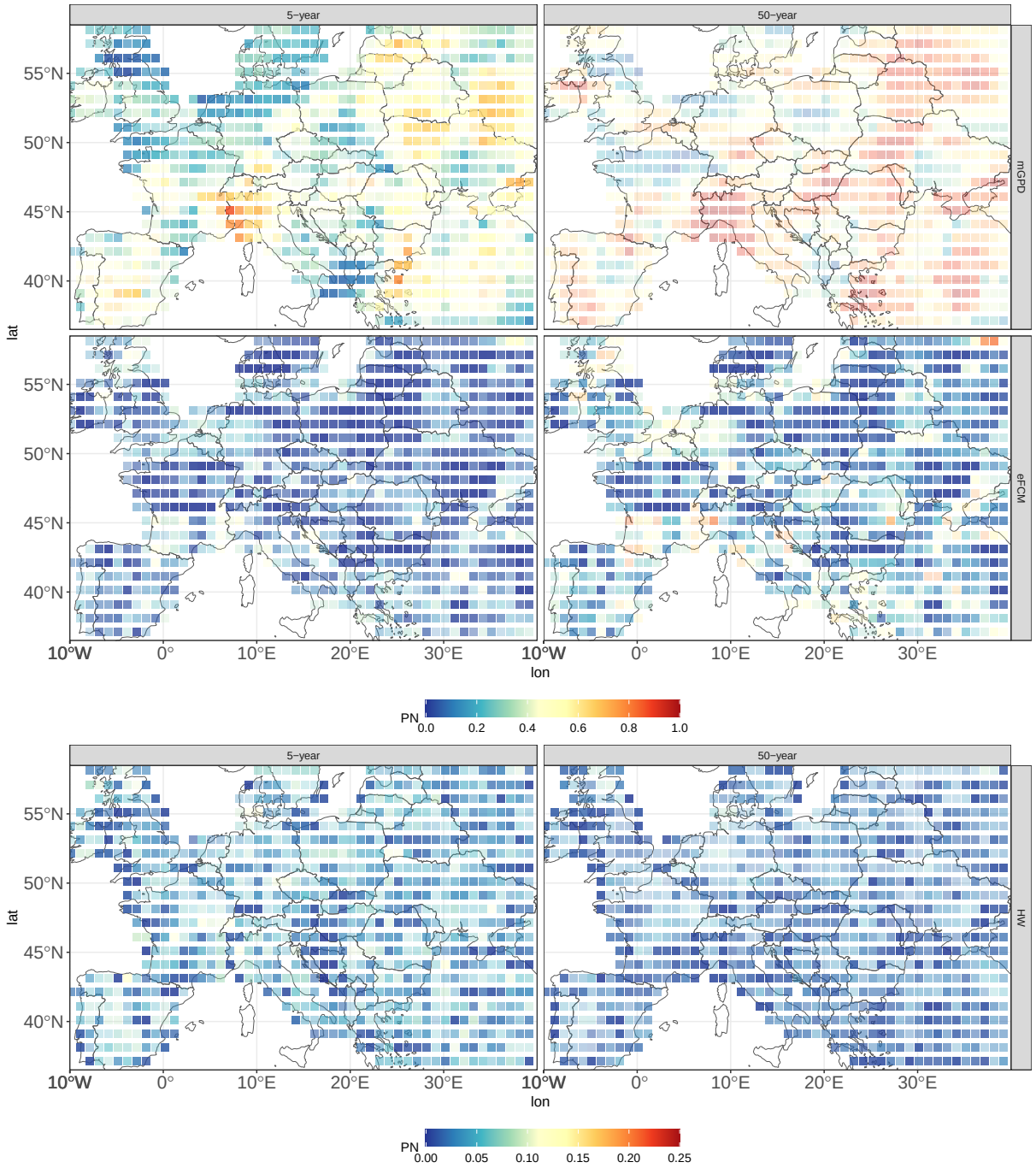

Figure A.4: Spatial probability of necessary causation $\mathrm{PN} = \max(1 - p_0/p_1, 0)$, with $p_t = \Pr\{\boldsymbol{w}^\top \boldsymbol{X}^{(t)} > v\}$ from (4.1), for extreme precipitation across Europe. Columns use 5-year and 50-year return levels for v; rows show the mGPD, eFCM, and HW fits. Colour represents PN on the displayed $[0, 1]$ scales, and opacity represents confidence-interval width, with paler points indicating greater uncertainty.

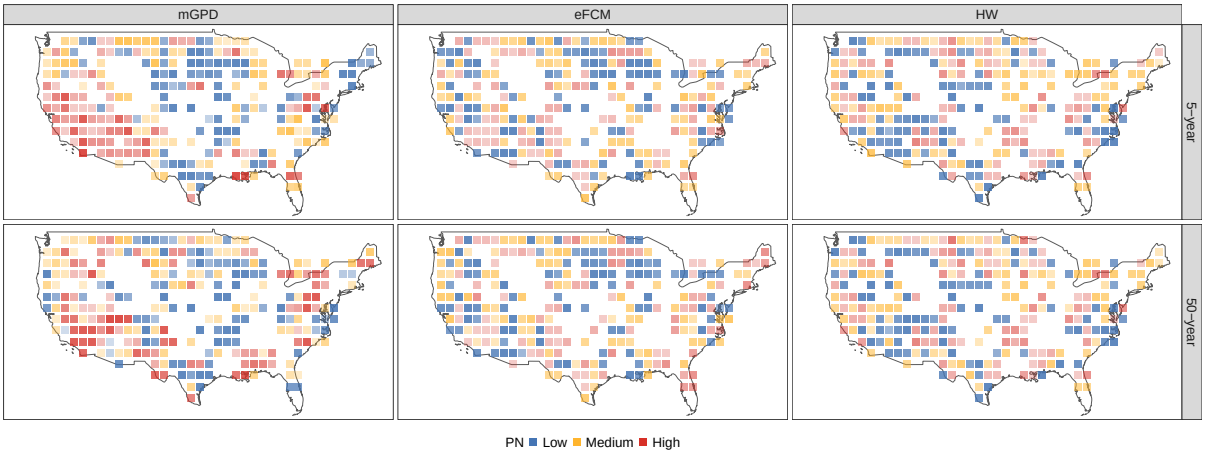

Figure A.5: Spatial probability of necessary causation $\mathrm{PN} = \max(1 - p_0/p_1, 0)$, with $p_t = \Pr\{\boldsymbol{w}^\top \boldsymbol{X}^{(t)} > v\}$ from (4.1), for extreme precipitation across the contiguous U.S. Rows use 5-year and 50-year return levels for v; columns show the eFCM, HW, and mGPD fits. PN is grouped into the displayed low, medium, and high categories, and opacity represents confidence-interval width, with paler points indicating greater uncertainty.

B Appendix material for Chapter 5

B.1 Proof of Theorem 5.3

The proof of asymptotic normality for the TIEE estimator $\hat{\theta}_{d,\tau}$ follows the standard approach for semi-parametric Z-estimators. We assume that the true parameters $\theta_{d,\tau}$ and ζ are fixed, and we emphasise the dependence of the empirical functional $\Phi_{d,n}$ of the estimator of ζ by writing

$$\Phi_{d,n}(\theta, \hat{\zeta}) = \frac{1}{n} \sum_{i=1}^{n} g_{\text{TIEE},d}(W(i), \tau, \theta, \hat{\zeta}).$$

As we know, the estimator $\hat{\theta}_{d,\tau}$ is defined as the root of the empirical estimating equation $\Phi_{d,n}(\hat{\theta}_{d,\tau}, \hat{\zeta}) = 0$. Applying the mean value theorem to $\Phi_{d,n}(\theta, \hat{\zeta})$ in θ around the true parameter $\theta_{d,\tau}$ yields

$$0 = \Phi_{d,n}(\hat{\theta}_{d,\tau}, \hat{\zeta}) = \Phi_{d,n}(\theta_{d,\tau}, \hat{\zeta}) + (\hat{\theta}_{d,\tau} - \theta_{d,\tau}) \cdot \frac{\partial}{\partial \theta} \Phi_{d,n}(\tilde{\theta}, \hat{\zeta}),$$

where $\tilde{\theta}$ is an intermediate value between $\hat{\theta}_{d,\tau}$ and $\theta_{d,\tau}$. Rearranging terms, we get

$$\sqrt{n}(\hat{\theta}_{d,\tau} - \theta_{d,\tau}) = -\frac{\sqrt{n}\Phi_{d,n}(\theta_{d,\tau}, \hat{\zeta})}{\frac{\partial}{\partial \theta}\Phi_{d,n}(\tilde{\theta}, \hat{\zeta})}. \tag{B.1}$$

By the consistency of $\hat{\theta}_{d,\tau}$ established in Theorem 5.2, we have $\tilde{\theta} \xrightarrow{P} \theta_{d,\tau}$ and by Assumption 5.4, $\hat{\zeta} \xrightarrow{P} \zeta$. Furthermore, Assumption 5.5(1) ensures that the partial derivative $\frac{\partial}{\partial \theta}\Phi_d(\theta, \zeta)$ is continuous in a neighbourhood of $\theta_{d,\tau}$. The uniform convergence of the empirical derivative to its population counterpart,

$$\frac{\partial}{\partial \theta}\Phi_{d,n}(\theta, \zeta) \xrightarrow{P} \frac{\partial}{\partial \theta}\Phi_d(\theta, \zeta),$$

combined with the consistency of the estimated parameters, ensures that

$$\frac{\partial}{\partial \theta}\Phi_{d,n}(\tilde{\theta}, \hat{\zeta}) \xrightarrow{P} \frac{\partial}{\partial \theta}\Phi_d(\theta_{d,\tau}, \zeta). \tag{B.2}$$

Since Assumption 5.5(1) requires $\frac{\partial}{\partial\theta}\Phi_d(\theta_{d,\tau}) \neq 0$, the denominator in (B.1) converges to a non-zero constant, ensuring the asymptotic normality result is well-defined.

Now we turn our attention to the numerator in (B.1). Since this is a function of the estimated nuisance parameter $\hat{\zeta}$, we exploit the pathwise differentiability of the moment functional with respect to ζ (Assumption 5.5(2)) to obtain the following asymptotic linear expansion around the true parameter ζ:

$$\sqrt{n}\Phi_{d,n}(\theta_{d,\tau}, \hat{\zeta}) = \sqrt{n}\Phi_{d,n}(\theta_{d,\tau}, \zeta) + \sqrt{n}(\hat{\zeta} - \zeta)^\top \cdot \nabla_\zeta \Phi_d(\theta_{d,\tau}, \zeta) + o_p(1), \tag{B.3}$$

where $\nabla_\zeta \Phi_d(\theta_{d,\tau}, \zeta)$ is the gradient of the population functional with respect to ζ, evaluated at the true parameters. Since $\Phi_d(\theta_{d,\tau}, \zeta) = 0$ is the moment condition identifying the true quantile $\theta_{d,\tau}$, the moment condition is locally unbiased with respect to ζ at the root (this is a standard property for efficient semi-parametric estimators). This condition implies that the term involving the estimation error of $\hat{\zeta}$ vanishes asymptotically. Therefore, the numerator in (B.3) simplifies to

$$\sqrt{n}\Phi_{d,n}(\theta_{d,\tau}, \hat{\zeta}) = \sqrt{n}\Phi_{d,n}(\theta_{d,\tau}, \zeta) + o_p(1).$$

Substituting the definition of $\Phi_{d,n}(\theta_{d,\tau}, \zeta)$ and using the identification condition

$$\mathbb{E}[g_{\text{TIEE},d}(W(i), \tau, \theta_{d,\tau}, \zeta)] = \Phi_d(\theta_{d,\tau}, \zeta) = 0,$$

we obtain

$$\sqrt{n}\Phi_{d,n}(\theta_{d,\tau}, \hat{\zeta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{\text{TIEE},d}(W_i, \tau, \theta_{d,\tau}, \zeta) + o_p(1),$$

which, by the central limit theorem, converges to a mean-zero Gaussian distribution with finite variance $\sigma^2_{d,\tau} = \text{Var}(g_{\text{TIEE},d}(W_i, \tau, \theta_{d,\tau}, \zeta))$. Combining (via Slutsky's theorem) this convergence result with that of (B.2), we have

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{d,\tau} - \theta_{d,\tau}) &= -\frac{\sqrt{n}\Phi_{d,n}(\theta_{d,\tau}, \hat{\zeta})}{\frac{\partial}{\partial\theta}\Phi_{d,n}(\tilde{\theta}, \hat{\zeta})} \\ &\xrightarrow{d} -\frac{\mathcal{N}(0, \sigma_q^2)}{\frac{\partial}{\partial\theta}\Phi_d(\theta_{d,\tau}, \zeta)} \equiv \mathcal{N}\left(0, \frac{\sigma_q^2}{[\frac{\partial}{\partial\theta}\Phi_d(\theta_{d,\tau}, \zeta)]^2}\right).\end{aligned}$$

B.2 Variance under non-vanishing nuisance sensitivity

B.2.1 Derivation of the variance formula

The standard derivation in Theorem 5.3 assumes that the estimation of the nuisance parameter ζ does not affect the asymptotic distribution of $\theta_{d,\tau}$ through the orthogonality condition $\nabla_\zeta \Phi_d(\theta_{d,\tau}, \zeta) = 0$. When this condition fails to hold, we must incorporate the estimation uncertainty of $\hat{\zeta}$ into the calculation of the asymptotic variance $V_{d,\tau}$. This section derives the general variance formula that accounts for the influence of nuisance parameter estimation.

Let ζ be a $k \times 1$ vector of nuisance parameters. The estimators $(\hat{\theta}_{d,\tau}, \hat{\zeta})$ are obtained by solving the empirical moment conditions

$$\Phi_{d,n}(\hat{\theta}_{d,\tau}, \hat{\zeta}) = \frac{1}{n}\sum_{i=1}^{n} g_{\text{TIEE},d}(W_i, \hat{\theta}_{d,\tau}, \hat{\zeta}) = 0, \quad \text{and} \quad \Psi_{d,n}(\hat{\zeta}) = \frac{1}{n}\sum_{i=1}^{n} \psi(W_i, \hat{\zeta}) = 0,$$

where $\psi(W_i, \hat{\zeta})$ is an estimating function defining the nuisance parameter estimator ζ through the moment condition $\mathbb{E}[\psi(W_i, \hat{\zeta})] = 0$. Let the stacked moment function m and its average be defined as

$$m(W, \theta, \zeta) = \begin{pmatrix} g_{\text{TIEE},d}(W, \theta, \zeta) \\ \psi(W, \zeta) \end{pmatrix}, \qquad \bar{m}_n(\theta, \zeta) = \frac{1}{n}\sum_{i=1}^{n} m(W_i, \theta, \zeta).$$

Under standard regularity conditions, the joint asymptotic distribution of $(\hat{\theta}_{d,\tau}, \hat{\zeta})$ is

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_{d,\tau} - \theta_{d,\tau} \\ \hat{\zeta} - \zeta \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, V_{\text{full}}),$$

where V_{full} is the sandwich covariance matrix given by $V_{\text{full}} = A^{-1}B(A^{-1})^\top$, where A is the Jacobian matrix and B is the covariance matrix. Specifically, A is the expected gradient of the moment function m with respect to the parameters (θ, ζ), i.e.,

$$A = \mathbb{E}\left[\frac{\partial m(W, \theta_{d,\tau}, \zeta)}{\partial(\theta, \zeta)}\right] = \begin{pmatrix} \mathbb{E}[\partial_\theta g_{\text{TIEE},d}] & \mathbb{E}[\partial_\zeta g_{\text{TIEE},d}] \\ \mathbb{E}[\partial_\theta \psi] & \mathbb{E}[\partial_\zeta \psi] \end{pmatrix}.$$

Under the natural assumption that the estimating equation for ζ does not depend on θ, we have $\mathbb{E}[\partial_\theta \psi] = 0$. Therefore, the Jacobian matrix takes the block upper-triangular form

$$A = \begin{pmatrix} \Phi'(\theta_{d,\tau}) & A_\zeta \\ 0 & M \end{pmatrix}, \tag{B.4}$$

where $\Phi'(\theta_{d,\tau}) = \mathbb{E}[\partial_\theta g_{\text{TIEE},d}(W, \theta_{d,\tau}, \zeta)]$ is a scalar representing the sensitivity of the moment condition to changes in θ; $A_\zeta = \mathbb{E}[\partial_\zeta g_{\text{TIEE},d}(W, \theta_{d,\tau}, \zeta)] = \nabla_\zeta \Phi_d(\theta_{d,\tau}, \zeta)$ is a $1 \times k$ row vector capturing the sensitivity to nuisance parameters; and $M = \mathbb{E}[\partial_\zeta \psi(W, \zeta)]$ is a $k \times k$ matrix governing the estimation of ζ.

The matrix B is the covariance matrix of the stacked moment functions evaluated at the true parameters, i.e.,

$$B = \text{Var}(m(W, \theta_{d,\tau}, \zeta)) = \begin{pmatrix} \text{Var}(g_{\text{TIEE},d}) & \text{Cov}(g_{\text{TIEE},d}, \psi) \\ \text{Cov}(\psi, g_{\text{TIEE},d}) & \text{Var}(\psi) \end{pmatrix} = \begin{pmatrix} \sigma^2_{d,\tau} & \Sigma_{g,\psi} \\ \Sigma^\top_{g,\psi} & \Sigma_\psi \end{pmatrix},$$

where $\sigma^2_{d,\tau} = \text{Var}(g_{\text{TIEE},d}(W, \theta_{d,\tau}, \zeta))$ is the scalar variance of the moment function for $\theta_{d,\tau}$; $\Sigma_\psi = \text{Var}(\psi(W, \zeta))$ is the $k \times k$ covariance matrix of the moment functions for ζ; and $\Sigma_{g,\psi} = \text{Cov}(g_{\text{TIEE},d}(W, \theta_{d,\tau}, \zeta), \psi(W, \zeta))$ is the $1 \times k$ row vector of covariances between the two sets of moment functions.

$V_{d,\tau}$ is defined by the $(1,1)$ block of V_{full}, i.e., $V_{d,\tau} = \text{Var}(\sqrt{n}(\hat{\theta}_{d,\tau} - \theta_{d,\tau}))$. For a block upper-triangular matrix of the form given in (B.4), the inverse is

$$A^{-1} = \begin{pmatrix} (\Phi')^{-1} & -(\Phi')^{-1} A_\zeta M^{-1} \\ 0 & M^{-1} \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix}. \tag{B.5}$$

$V_{d,\tau}$ can then be computed as

$$V_{d,\tau} = U_{11} B_{11} U_{11}^\top + U_{12} B_{21} U_{11}^\top + U_{11} B_{12} U_{12}^\top + U_{12} B_{22} U_{12}^\top, \tag{B.6}$$

where $B_{11} = \sigma^2_{d,\tau}$, $B_{12} = \Sigma_{g,\psi}$, $B_{21} = \Sigma^\top_{g,\psi}$, and $B_{22} = \Sigma_\psi$. Since $V_{d,\tau}$ is a scalar quantity, the second and third terms in (B.6) are transposes of each other and therefore equal. Substituting the expressions for U_{11} and U_{12} from (B.5), we obtain the general variance formula

$$V_q = \frac{\sigma^2_q + A_\zeta V_\zeta A^\top_\zeta - 2A_\zeta C}{[\Phi'(\theta_{d,\tau})]^2}, \tag{B.7}$$

where $V_\zeta = M^{-1}\Sigma_\psi(M^{-1})^\top$ and $C = M^{-1}\Sigma^\top_{g,\psi}$. Here, V_ζ represents the asymptotic variance of $\hat{\zeta}$, and C captures the scaled covariance between the moment functions $\overline{g}$ and ψ. The formula in (B.7) contains three components: $\sigma^2_{d,\tau}$, the intrinsic variability of the moment condition for $\theta_{d,\tau}$; $A_\zeta V_\zeta A^\top_\zeta$, the variance inflation due to the estimation of nuisance parameters; and $-2A_\zeta C$, a correction term accounting for the correlation between the estimation procedures for $\theta_{d,\tau}$ and ζ.

B.2.2 Consistent plug-in estimation

A consistent plug-in estimator $\hat{V}_{d,\tau}$ can be constructed by replacing all population quantities in (B.7) with their empirical counterparts, i.e.,

$$\hat{V}_q = \frac{\hat{\sigma}^2_{d,\tau} + \hat{A}_\zeta \hat{V}_\zeta \hat{A}^\top_\zeta - 2\hat{A}_\zeta \hat{C}}{(\hat{\Phi}')^2},$$

where the empirical quantities are defined as:

$$\begin{aligned}
\hat{\Phi}' &= \frac{1}{n}\sum_{i=1}^n \partial_\theta g_{\text{TIEE},d}(W_i, \hat{\theta}_q, \hat{\zeta}),\\
\hat{A}_\zeta &= \frac{1}{n}\sum_{i=1}^n \partial_\zeta g_{\text{TIEE},d}(W_i, \hat{\theta}_q, \hat{\zeta}),\\
\hat{M} &= \frac{1}{n}\sum_{i=1}^n \partial_\zeta \psi(W_i; \hat{\zeta}),\\
\hat{V}_\zeta &= \hat{M}^{-1}\hat{\Sigma}_\psi(\hat{M}^{-1})^\top,\\
\hat{C} &= \hat{M}^{-1}\hat{\Sigma}^\top_{g,\psi}.
\end{aligned}$$

The empirical second moments $\hat{\sigma}^2_{d,\tau}$, $\hat{\Sigma}_\psi$, and $\hat{\Sigma}_{g,\psi}$ are computed from the centred moment functions as

$$\hat{\sigma}^2_{d,\tau} = \frac{1}{n}\sum_{i=1}^{n}[g_{\text{TIEE},d}(W_i, \hat{\theta}_{d,\tau}, \hat{\zeta})]^2,$$
$$\hat{\Sigma}_\psi = \frac{1}{n}\sum_{i=1}^{n}\psi(W_i, \hat{\zeta})\psi(W_i, \hat{\zeta})^\top,$$
$$\hat{\Sigma}_{g,\psi} = \frac{1}{n}\sum_{i=1}^{n}g_{\text{TIEE},d}(W_i, \hat{\theta}_{d,\tau}, \hat{\zeta})\psi(W_i, \hat{\zeta})^\top.$$

Under standard regularity conditions, $\hat{V}_{d,\tau} \xrightarrow{P} V_{d,\tau}$, ensuring consistent variance estimation for constructing confidence intervals and hypothesis tests.

B.2.3 Special cases

The general variance formula in (B.7) encompasses two important special cases that arise in practice.

Case 1: Orthogonality. If the orthogonality condition $\nabla_\zeta \Phi_d(\theta_{d,\tau}, \zeta) = 0$ holds, as assumed in Theorem 5.3, then $A_\zeta = 0$. Consequently, both the variance inflation term $A_\zeta V_\zeta A_\zeta^\top$ and the cross-term $2A_\zeta C$ vanish, and the variance simplifies to

$$V_{d,\tau} = \frac{\sigma_q^2}{[\Phi'(\theta_{d,\tau})]^2},$$

recovering the simpler result stated in Theorem 5.3. This is the most efficient case, as the nuisance parameter estimation introduces no additional variability.

Case 2: Uncorrelatedness. If $\Sigma_{g,\psi} = 0$ then $C = 0$ and the cross-term $2A_\zeta C$ disappears. This case can arise, for example, through sample-splitting or cross-fitting procedures that empirically decouple the estimation of $g_{\text{TIEE},d}$ and ψ. In this case, the variance becomes

$$V_{d,\tau} = \frac{\sigma_q^2 + A_\zeta V_\zeta A_\zeta^\top}{[\Phi'(\theta_{d,\tau})]^2}.$$

This case exhibits only the variance inflation due to nuisance parameter estimation, without the potentially beneficial correction from correlated estimation errors.

C Appendix material for Chapter 6

C.1 Additional high-dimensional scaling metrics

For completeness, Figure C.1 reports the corresponding Precision and Recall boxplots for the high-dimensional scaling experiment from Section 6.5. These metrics complement the main-text F_1 and SHD summaries by showing how screening sensitivity and positive predictive value change with dimension across methods.

C.2 Proof of Theorem 6.1

Proof. We prove the forward and backward bounds separately. Throughout the proof, $Y_j = 1/(1 - F_j(X_j))$ denotes the Pareto-standardised variable, $Z_j = \log Y_j - \log u$ the log-tail coordinate, and $\tilde{u}$ the X-scale threshold corresponding to u via the probability integral transform, so that $\{X_j > \tilde{u}\} \equiv \{Y_j > u\} \equiv \{Z_j > 0\}$. All conditioning events are equivalent up to monotone transformations.

Forward direction ($1 \to 2$). Condition on the tail event $\{Z_1 > 0\}$, which is equivalent to $\{Y_1 > u\}$ and hence to $\{X_1 > \tilde{u}\}$ for a corresponding threshold $\tilde{u} \to \infty$ under the exact probability integral transform. Under the max-linear specification,

$$X_2 = \max(cX_1, \epsilon_2).$$

Since ϵ_2 is independent of X_1 and $\epsilon_2 \sim$ Fréchet(1), we have, for $\tilde{u}$ sufficiently large,

$$\begin{aligned} \Pr(\epsilon_2 > cX_1 \mid X_1 > \tilde{u}) &= \mathbb{E}\big[\Pr(\epsilon_2 > cX_1 \mid X_1) \,\big|\, X_1 > \tilde{u}\big] \\ &= \mathbb{E}\left[1 - \exp\left(-\frac{1}{cX_1}\right) \,\middle|\, X_1 > \tilde{u}\right] \\ &\le \mathbb{E}\left[\frac{1}{cX_1} \,\middle|\, X_1 > \tilde{u}\right] \le \frac{1}{c\tilde{u}} \;\longrightarrow\; 0, \end{aligned} \tag{C.1}$$

as $\tilde{u} \to \infty$. Therefore,

$$\Pr(X_2 = cX_1 \mid X_1 > \tilde{u}) \;\to\; 1.$$

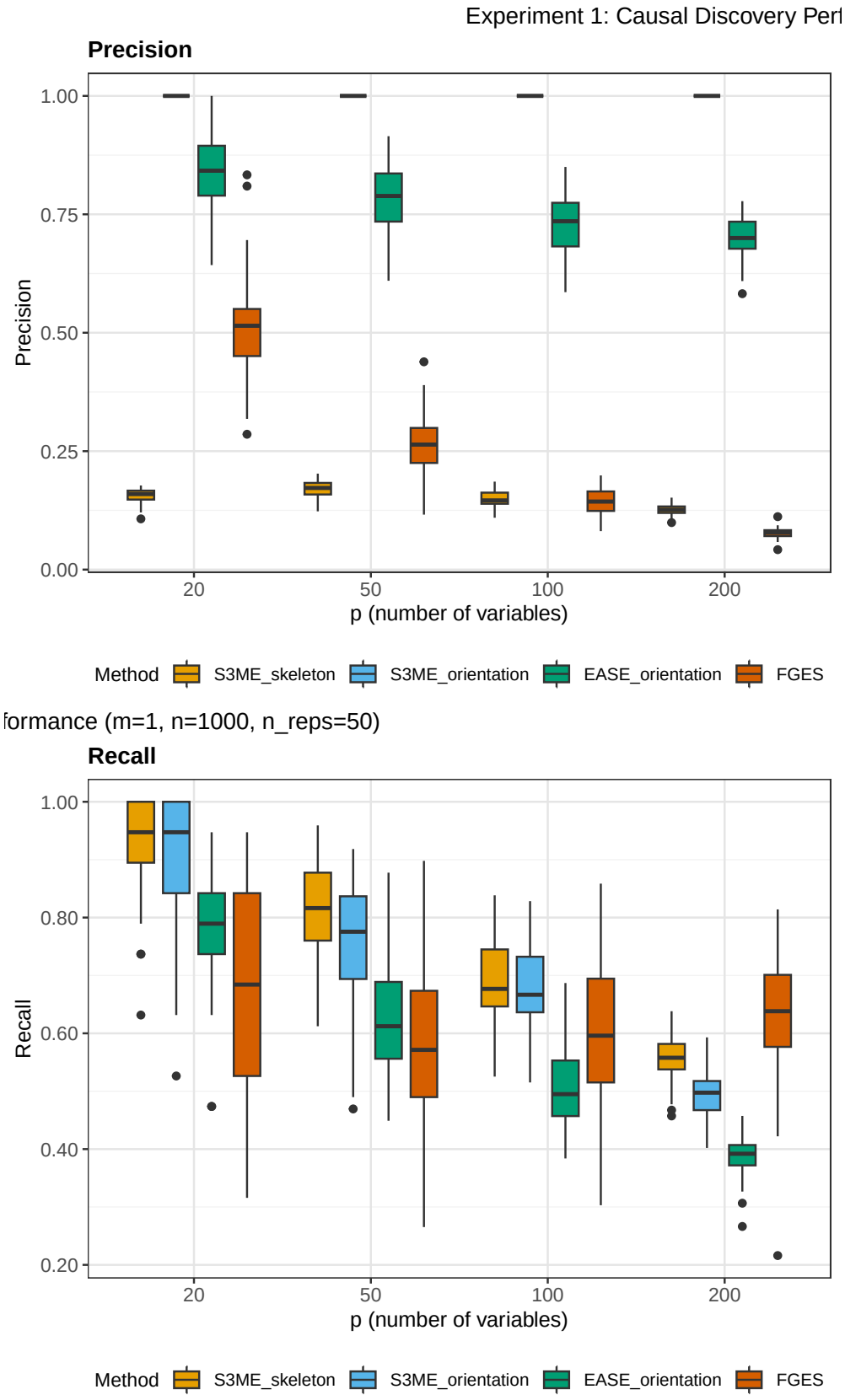

Figure C.1: High-dimensional scaling results for $m = 1$, $n = 1000$, and 50 Monte Carlo replicates. The upper and lower panels show precision and recall, respectively, across $p \in \{20, 50, 100, 200\}$ for the S3ME skeleton, EASE, FGES, and the full S3ME procedure. Precision is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, and recall is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, using the directed or undirected evaluation domain associated with the reported stage.

Since $\Pr(X_2 = cX_1 \mid X_1 > \tilde{u}) \to 1$, we have $X_2/X_1 \to c$ in probability on $\{X_1 > \tilde{u}\}$. It remains to track the effect of the Pareto standardization $Y_j = 1/(1 - F_j(X_j))$. For $X_1 \sim$ Fréchet(1), $1 - F_1(x) \sim 1/x$ as $x \to \infty$, so $Y_1 \sim X_1$ uniformly on $\{X_1 > \tilde{u}\}$. For $X_2 = \max(cX_1, \epsilon_2)$, $\Pr(X_2 > x) \sim (c+1)/x$ and hence $1-F_2(x) \sim (c+1)/x$, so $Y_2 \sim X_2/(c+1)$ uniformly on $\{X_2 > \tilde{u}\}$. Combining these with $\Pr(X_2 = cX_1 \mid X_1 > \tilde{u}) \to 1$ and by standard arguments for conditional convergence in probability,

$$\frac{Y_2}{Y_1} \;\to\; \frac{c}{c+1} \;=:\; \kappa \qquad \text{in probability given } X_1 > \tilde{u}.$$

The log-tail coordinates $Z_j = \log Y_j - \log u$ then satisfy

$$Z_2 - Z_1 \to \log \kappa \quad \text{in probability given } Z_1 > 0.$$

Choosing any finite $c_0 < \log \kappa$ and the predictor $\hat{Z}_2 = \max(Z_1 + \log \kappa,\, c_0) = \max(Z_1,\, c_0 - \log \kappa) + \log \kappa$ yields $\hat{Z}_2 \in \mathcal{H}$ and $\hat{Z}_2 = Z_1 + \log \kappa$ on the event $\{Z_1 > 0\}$. Hence,

$$\mathbb{E}\big[|Z_2 - \hat{Z}_2| \mid Z_1 > 0\big] \;\to\; 0,$$

so the optimal tail prediction risk satisfies $R^{\star}_{2|1}(u) \to 0$ as $u \to \infty$.

Backward direction ($2 \to 1$). Now condition on $\{X_2 > \tilde{u}\}$ for $\tilde{u} \to \infty$. The exceedance $\{X_2 > \tilde{u}\}$ can occur via two asymptotically distinct mechanisms:

(i) propagation from an upstream extreme cX_1; or (ii) a local extreme innovation ϵ_2 dominating cX_1.

Specifically, since X_1 and ϵ_2 are independent Fréchet(1) variables,

$$\Pr(\epsilon_2 > u) \sim u^{-1}, \qquad \Pr(cX_1 > u) = \Pr(X_1 > u/c) \sim c\, u^{-1}.$$

By independence, the intersection term is $o(u^{-1})$, so

$$\Pr(X_2 > u) = \Pr(\max(cX_1, \epsilon_2) > u) \sim (c+1)\, u^{-1}.$$

Meanwhile, on the event $\{\epsilon_2 > cX_1,\ X_2 > u\}$, we have $X_2 = \epsilon_2$, and

$$\Pr(\epsilon_2 > cX_1,\ X_2 > u) = \Pr(\epsilon_2 > u,\ \epsilon_2 > cX_1) \sim \Pr(\epsilon_2 > u) \sim u^{-1}.$$

Write $A_u := \{X_2 > \tilde{u}\}$ and $B := \{\epsilon_2 > cX_1\}$. The calculation above gives

$$\lim_{u\to\infty} \Pr(B \mid A_u) = \frac{1}{c+1}. \tag{C.2}$$

Since node 1 is a source, $X_1 = \epsilon_1$, so on $A_u \cap B$ we have $X_2 = \epsilon_2$ and $\epsilon_1 < \epsilon_2/c$. Any predictor $h \in \mathcal{H}$ can be written as $h(z) = \max(z, \gamma) + d$; equivalently, setting $c_{2\to1} = d$ and $c_0' = \gamma + d$ gives the reparametrization $h(z) = \max(z + c_{2\to1}, c_0')$, which we use below. For any max-linear envelope predictor $\hat{Z}_1 = \max(Z_2 + c_{2\to1}, c_0')$, by the law of total expectation over B and B^c within A_u, and dropping the non-negative B^c term, we have $\mathbb{E}[|Z_1 - \hat{Z}_1| \mid A_u] \geq \Pr(B \mid A_u)\,\mathbb{E}[|Z_1 - \hat{Z}_1| \mid A_u, B]$ for every predictor. Since $\Pr(B \mid A_u)$ is free of the predictor parameters, taking the infimum over $(c_{2\to1}, c_0')$ on both sides yields

$$\inf_{c_{2\to1},\,c_0'} \mathbb{E}\big[|Z_1 - \hat{Z}_1| \mid A_u\big] \ \geq\ \Pr(B \mid A_u) \inf_{c_{2\to1},\,c_0'} \mathbb{E}\big[|Z_1 - \max(Z_2 + c_{2\to1}, c_0')| \mid A_u, B\big]\,.$$

We now lower-bound the inner infimum on $A_u \cap B$. On this event, $X_2 = \epsilon_2$ and $X_1 = \epsilon_1$ with $\epsilon_1 \perp \epsilon_2$. Given $\epsilon_2 = y$ and $\epsilon_2 > c\epsilon_1$, the independence of ϵ_1 and ϵ_2 yields, for $x < y/c$,

$$\Pr(\epsilon_1 \leq x \mid \epsilon_2 = y,\ \epsilon_2 > c\epsilon_1) = \frac{\Pr(\epsilon_1 \leq x,\ \epsilon_1 < y/c)}{\Pr(\epsilon_1 < y/c)} = \frac{F_1(x)}{F_1(y/c)}.$$

Since $y \asymp ue^z \to \infty$ on $\{Z_2 = z,\ A_u \cap B\}$, we have $F_1(y/c) \to 1$, and hence for every fixed x,

$$\Pr(\epsilon_1 \leq x \mid Z_2 = z,\ A_u \cap B) \to F_1(x).$$

That is, the conditional distribution of ϵ_1 given (Z_2, A_u, B) converges pointwise to the unconditional Fréchet(1), for each fixed x, as $u \to \infty$.

Since $Y_1 = 1/(1-F_1(\epsilon_1))$ is the exact probability-integral-transform of ϵ_1 to unit Pareto, $\log Y_1 \sim \mathrm{Exp}(1)$ exactly, and $Z_1 = \log Y_1 - \log u$. By the convergence above, for any $h \in \mathcal{H}$ and any z,

$$\mathbb{E}[|Z_1 - h(z)| \mid Z_2 = z,\, A_u \cap B] = \mathbb{E}[|\log Y_1 - \log u - h(z)|] + o(1) \;\geq\; \mathbb{E}[|\log Y_1 - \log 2|] + o(1),$$

where the inequality uses the fact that $\log 2 - \log u$ is the median of $Z_1 = \log Y_1 - \log u$, and the median minimizes the ℓ_1 risk. Since $\mathbb{E}[|\mathrm{Exp}(1) - \log 2|] = \log 2$, we have

$$\liminf_{u\to\infty} \inf_{h\in\mathcal{H}} \mathbb{E}[|Z_1 - h(Z_2)| \mid A_u \cap B] \;\geq\; \log 2 \;>\; 0.$$

Combining the two bounds gives

$$\liminf_{u\to\infty} R^{\star}_{1|2}(u) \geq \left(\lim_{u\to\infty} \Pr(B \mid A_u)\right)\left(\liminf_{u\to\infty} \inf_{h\in\mathcal{H}} \mathbb{E}[|Z_1 - h(Z_2)| \mid A_u \cap B]\right) \geq \frac{\log 2}{c+1} > 0.$$

From tail risk separation to identifiability. Finally, for the ℓ_1-based max-linear envelope scoring criterion, the nodewise contribution is monotone in the (tail-weighted) absolute prediction error. Since the forward tail risk converges to zero while the reverse tail risk has a strictly positive lim inf as $u \to \infty$, any monotone ℓ_1-based score evaluated at a sufficiently large threshold u strictly prefers the true direction $1 \to 2$ over $2 \to 1$, yielding asymptotic identifiability. □

C.3 Nodewise score separation and global identifiability

This section provides the technical verification that the abstract population-separation conditions of Section 6.2.3 are satisfied by the specific EBIC-based score used in our procedure.

Proposition C.1 (Superset non-improvement under the max-linear benchmark)**.** *Under the fully observed recursive max-linear specification of Assumption 6.1 with* $f_j(\{x_m\}) = \max_{m\in \mathrm{pa}(j)} a_{mj} x_m$*, unit Fréchet innovations, and no latent confounders, for any fixed node* j *and any superset* $A \supseteq \mathrm{pa}^*(j)$*,*

$$R^*_{j|A}(u) = R^*_{j|\mathrm{pa}^*(j)}(u).$$

In particular, adding spurious parents does not improve population tail risk.

Proof. Fix node j and write $A = \mathrm{pa}^*(j) \cup B$ with $B \cap \mathrm{pa}^*(j) = \emptyset$. Under the recursive max-linear model and Proposition C.2, compare the log-sum-exp proxies under the true parent set and superset:

$$T_0 := \sum_{m \in \mathrm{pa}^*(j)} a_{mj} e^{L_m} + \epsilon_j, \qquad T_B(c_B) := \sum_{m \in B} e^{c_{m\to j} + L_m},$$

$$\tilde{L}_j^A(c_B) = \log\big(T_0 + T_B(c_B)\big), \qquad \tilde{L}_j^{\mathrm{pa}} = \log(T_0).$$

Since $T_B(c_B) \geq 0$, we have $\tilde{L}_j^A(c_B) \geq \tilde{L}_j^{\mathrm{pa}} \geq L_j$ pointwise, so

$$|L_j - \tilde{L}_j^A(c_B)| \geq |L_j - \tilde{L}_j^{\mathrm{pa}}|.$$

Taking expectations and infimum over finite c_B yields

$$R^*_{j|A}(u) \geq R^*_{j|\mathrm{pa}^*(j)}(u).$$

Conversely, set $c^{(r)}_{m\to j} = -r$ for $m \in B$. Then $T_B(c_B^{(r)}) \downarrow 0$ and $\tilde{L}_j^A(c_B^{(r)}) \downarrow \tilde{L}_j^{\mathrm{pa}}$ pointwise as $r \to \infty$. By monotone convergence of nonnegative errors,

$$\lim_{r\to\infty} R_{j|A}(u; c_B^{(r)}) = R^*_{j|\mathrm{pa}^*(j)}(u),$$

which implies $R^*_{j|A}(u) \leq R^*_{j|\mathrm{pa}^*(j)}(u)$. Therefore equality holds. □

Proof of Theorem 6.2. Fix $A \subseteq \mathrm{nb}^*(j)$ with $A \neq \mathrm{pa}^*(j)$. There are two cases.

Case 1: $A \not\supseteq \mathrm{pa}^*(j)$. By Assumption 6.2, there exist constants $\Delta_{\mathrm{miss},j} > 0$ and $u_{0,j}$ such that for all $u \geq u_{0,j}$,

$$R^*_{j|A}(u) - R^*_{j|\mathrm{pa}^*(j)}(u) \geq \Delta_{\mathrm{miss},j} > 0.$$

Hence any parent set omitting at least one true parent has strictly larger population risk.

Case 2: $A \supsetneq \mathrm{pa}^*(j)$. By Assumption 6.3,

$$R^*_{j|A}(u) \geq R^*_{j|\mathrm{pa}^*(j)}(u)$$

for sufficiently large u.

Thus $\mathrm{pa}^*(j)$ is a minimizer of $R^*_{j|A}(u)$. Moreover, any minimizer cannot be in Case 1, so every minimizer must contain $\mathrm{pa}^*(j)$. Therefore $\mathrm{pa}^*(j)$ is the unique inclusion-minimal minimizer. □

Proof of Corollary 6.1. For any skeleton-compatible DAG G, Theorem 6.2 gives nodewise inequalities

$$R^*_{j|\mathrm{pa}_G(j)}(u) \geq R^*_{j|\mathrm{pa}^*(j)}(u), \qquad j = 1, \ldots, p.$$

Summing over j yields

$$R^*(G; u) \geq R^*(\mathcal{G}^*; u),$$

so $\mathcal{G}^*$ is a global minimizer. If G is any global minimizer, equality must hold at every node, hence each $\mathrm{pa}_G(j)$ is a nodewise minimizer and must contain $\mathrm{pa}^*(j)$ by Theorem 6.2. Therefore every global minimizer contains $\mathcal{G}^*$ edgewise, and $\mathcal{G}^*$ is the unique inclusion-minimal global minimizer. □

C.4 Uniformly Bounded Approximation Error

Precise regularity conditions for Assumption 6.2. The compact statement of Assumption 6.2 in the main paper abbreviates the following rate conditions for the true undirected skeleton, the Hüsler–Reiss limit, and the proxy-adjusted model:

1. **(Sparsity and Intermediate Effective Sample Size)** For each node j, let $\mathcal{N}_j = \{i : \{i, j\} \in E^*_{\mathrm{skel}}\}$ and $s_j = |\mathcal{N}_j|$ denote its neighbor set and degree in the skeleton, and let $s_{\max} = \max_j s_j$. We assume $s_{\max} = o\left(\sqrt{\frac{k}{\log p}}\right)$.
2. **(Bounded Eigenvalues)** For each node j with true neighbour set $\mathcal{N}_j$, the submatrix $\Sigma_{\mathcal{N}_j\mathcal{N}_j}$ of the Hüsler–Reiss covariance matrix $\Sigma = \Theta^{-1}$ satisfies $\Lambda_{\min}(\Sigma_{\mathcal{N}_j\mathcal{N}_j}) \geq C_{\min}$ for some absolute constant $C_{\min} > 0$, where $\Lambda_{\min}(\cdot)$ denotes the smallest eigenvalue.
3. **(Minimum Signal Strength in the Tail)** The non-zero regression coefficients in the tail limit are sufficiently strong to be detected at effective sample size k. Specifically, for some sufficiently large constant $C > 0$, $\min_j \min_{i \in \mathcal{N}_j} |\beta^*_{ji}| \geq \frac{C}{\sqrt{C_{\min}}} \sqrt{\frac{s_{\max} \log p}{k}}$.

4. **(Score Bias Control)** For each nodewise regression after proxy partialling-out, the population score contribution of the log-sum-exp residual bias satisfies $\left\|\mathbb{E}[X^{\text{res}}\tilde{\Delta}]\right\|_\infty \leq c_\Delta \lambda$ for some constant $c_\Delta > 0$, where X^{res} denotes the residualized design and $\tilde{\Delta} = \Delta_{\text{model}} - \bar{\Delta}\mathbb{1}_k$ is the mean-centered approximation bias from replacing the max-linear structural equation by a log-sum-exp proxy, bounded by $\log(|\text{pa}(j)| + 1)$ per Proposition C.2 below.

Proposition C.2 (Uniformly bounded approximation error)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ follow a recursive max-linear SEM under Assumption 6.1 with $f_j(\{x_m\}) = \max_{m\in\text{pa}(j)} a_{mj}x_m$ (edge weights $a_{mj} > 0$) and unit Fréchet innovations. Write $\boldsymbol{L}_j = \log \boldsymbol{X}_j$. The structural equation for node $\boldsymbol{j}$ takes the form*

$$L_j \;=\; \bigvee_{m\in\text{pa}(j)} \big(L_m + \log a_{mj}\big) \;\vee\; \log \epsilon_j.$$

Define the Log-Sum-Exp approximation

$$\tilde{L}_j \;=\; \log\left(\sum_{m\in\text{pa}(j)} a_{mj}\, e^{L_m} + \epsilon_j\right),$$

and the approximation error $\mathcal{E}_j = \tilde{L}_j - L_j$. Then, for any observation, $\mathcal{E}_j$ is deterministically bounded:

$$0 \leq \mathcal{E}_j \leq \log\big(|\text{pa}(j)| + 1\big).$$

Consequently, the log-sum-exp proxy is equivalent to the max-linear structural equation up to a strictly bounded $O(1)$ constant, independent of the tail threshold $\boldsymbol{u}$.

Proof. Let $X_{(1)} = \max_{m\in\text{pa}(j)}(a_{mj}e^{L_m}) \vee \epsilon_j$ denote the maximum term in the true structural equation. The approximation error induced by the Log-Sum-Exp mapping is exactly

$$\mathcal{E}_j = \log\left(\sum_{m\in\text{pa}(j)} a_{mj}e^{L_m} + \epsilon_j\right) - \log X_{(1)} = \log\left(1 + \sum_{m\neq m^*} \frac{X_m}{X_{(1)}}\right),$$

where the sum inside the logarithm contains exactly $|\text{pa}(j)|$ non-maximum terms. By the definition of the maximum, the ratio $X_m/X_{(1)} \leq 1$ for all terms. Therefore, we deterministically obtain the uniform bound $0 \leq \mathcal{E}_j \leq \log(1 + |\text{pa}(j)|)$. □

C.5 Proof of Lemma 6.1

Proof of Lemma 6.1. By Theorem 6.3, condition on the event $E^*_{\text{skel}} \subseteq \hat{E}_{\text{skel}}$ with $|\hat{E}_{\text{skel}}| \le c_3 p s_{\max}$. For each node j, the parent set A must satisfy $A \subseteq \text{nb}(j; \hat{E}_{\text{skel}})$ and $|A| \le s_{\max}$. The number of such sets is at most

$$\sum_{m=0}^{s_{\max}} \binom{|\text{nb}(j; \hat{E}_{\text{skel}})|}{m} \le \sum_{m=0}^{s_{\max}} p^m \le (s_{\max}+1)\, p^{s_{\max}}.$$

Union-bounding the assumed per-set concentration over all A for fixed j,

$$\Pr\Big(\sup_A \big|\widehat{R}^{\text{dir}}_{j|A}(u) - R^*_{j|A}(u)\big| > \epsilon\Big) \le c_1 \exp\big(s_{\max}\log p + \log(s_{\max}+1) - c_2 k \min(\epsilon^2, \epsilon)\big).$$

A further union bound over p nodes gives

$$\Pr\Big(\max_{1\le j\le p} \sup_A \big|\widehat{R}^{\text{dir}}_{j|A}(u) - R^*_{j|A}(u)\big| > \epsilon\Big) \le c_1 \exp\big((s_{\max}+1)\log p + \log(s_{\max}+1) - c_2 k \min(\epsilon^2, \epsilon)\big).$$

Under $s_{\max} = o(\sqrt{k/\log p})$ and $\log p = o(k)$, the exponent diverges to $-\infty$ when $\epsilon = C\sqrt{(s_{\max}+1)\log p / k}$ for sufficiently large C, giving the stated rate. □

C.6 Proof of Theorem 6.3

Proof of Theorem 6.3. Fix a target node $j \in V$ and, to simplify notation, drop the subscript j. Because the proxy enters the nodewise regression as an unpenalized regressor, we first partial out its contribution and work with the residualized response and residualized design. To distinguish from the Pareto-standardized variables Y and the exceedance data $\{Z^{(i)}\}$ in Section 3, we write $\boldsymbol{y} \in \mathbb{R}^k$ for the resulting response vector of proxy-adjusted log-exceedances and $\boldsymbol{X} \in \mathbb{R}^{k\times(p-1)}$ for the corresponding design matrix of residualized predictors. Let β^* be the true sparse regression coefficient vector associated with the extremal precision matrix. Let $\mathcal{S} = \text{supp}(\beta^*)$ be the true support set and $|\mathcal{S}| = s$.

Under Assumption 4 and Proposition C.2, the proxy-adjusted tail-domain model can be written as $\boldsymbol{y} = \boldsymbol{X}\beta^* + W + \Delta_{\text{model}}$, where the stochastic noise W captures the Hüsler–Reiss approximation error and Δ_{model} is the log-sum-exp bias, deterministically bounded by $\log(|\text{pa}(j)|+1)$. By Proposition C.2, each component of Δ_{model} is bounded by $\log(|\text{pa}(j)|+$

1). The nodewise Lasso includes an unpenalized intercept, which absorbs the sample mean of Δ_{model} via the first-order KKT condition. The residual bias $\tilde{\Delta} = \Delta_{\text{model}} - \bar{\Delta}\,\mathbb{1}_k$ is therefore zero-mean and bounded. Under the extremal precision characterization, $\text{supp}(\beta^*)$ coincides with the true neighbourhood of node j.

The residual bias $\tilde{\Delta} = \Delta_{\text{model}} - \bar{\Delta}\mathbb{1}_k$ is centered, but its score contribution need not have zero population mean. We therefore write

$$\frac{1}{k}\boldsymbol{X}^\top\tilde{\Delta} = \mathbb{E}[X^{\text{res}}\tilde{\Delta}] + \left(\frac{1}{k}\boldsymbol{X}^\top\tilde{\Delta} - \mathbb{E}[X^{\text{res}}\tilde{\Delta}]\right).$$

By condition (4) of Assumption 8, $\|\mathbb{E}[X^{\text{res}}\tilde{\Delta}]\|_\infty \le c_\Delta\lambda$. For the fluctuation term, note that in the log-tail domain the coordinates are exponential-type; under Assumption 4, each coordinate of X^{res} is therefore sub-exponential (equivalently, has uniformly bounded ψ_1-Orlicz norm). Since $\tilde{\Delta}$ is bounded (Proposition C.2), the products

$$U_{t\ell} := X^{\text{res}}_{t\ell}\tilde{\Delta}_t - \mathbb{E}[X^{\text{res}}_{t\ell}\tilde{\Delta}_t]$$

are centered sub-exponential with $\|U_{t\ell}\|_{\psi_1} \le K_U$ uniformly in ℓ. Bernstein's inequality for sub-exponential sums (e.g., Wainwright, 2019, Corollary 2.8.3) yields, for any $x > 0$,

$$\Pr\left(\left|\frac{1}{k}\sum_{t=1}^{k} U_{t\ell}\right| > x\right) \le 2\exp\left[-ck\min\left(\frac{x^2}{K_U^2}, \frac{x}{K_U}\right)\right].$$

Applying a union bound over $\ell = 1, \dots, p$ and taking $x = C\sqrt{\log p/k}$, with $\log p = o(k)$ so that $x \to 0$ and the quadratic regime applies for large k, we obtain

$$\left\|\frac{1}{k}\boldsymbol{X}^\top\tilde{\Delta} - \mathbb{E}[X^{\text{res}}\tilde{\Delta}]\right\|_\infty = O_p\left(\sqrt{\frac{\log p}{k}}\right) = O_p(\lambda).$$

Hence $\|\frac{1}{k}\boldsymbol{X}^\top\tilde{\Delta}\|_\infty = O_p(\lambda)$, which is absorbed into the regularization parameter by enlarging the constant C in the choice of λ.

Retaining true edges (sure screening). We show that for each $i \in \mathcal{S}$ (each true neighbour), $\hat{\beta}_i \neq 0$ with high probability. By the standard ℓ_1 error bound for the Lasso under the restricted eigenvalue condition (Bühlmann and van de Geer, 2011), Assumption 8(2) gives

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{\sqrt{s}}{\kappa}\,\lambda \tag{C.3}$$

with probability at least $1 - c_5 \exp(-c_6 k\lambda^2)$, for a constant κ depending on $C_{\min}$ through the restricted eigenvalue condition (Assumption 8(2)). In particular,

$$\|\hat{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\|_\infty \leq \|\hat{\beta} - \beta^*\|_2 \leq \frac{\sqrt{s}}{\kappa}\,\lambda = O_p\left(\frac{1}{\sqrt{C_{\min}}}\sqrt{\frac{s_{\max}\log p}{k}}\right). \tag{C.4}$$

By Assumption 8(3), $\min_{i\in\mathcal{S}} |\beta^*_i| \geq \frac{C}{\sqrt{C_{\min}}}\sqrt{\frac{s_{\max}\log p}{k}}$ for a sufficiently large constant C, so $|\hat{\beta}_i - \beta^*_i| < |\beta^*_i|$ and hence $\hat{\beta}_i \neq 0$ for each $i \in \mathcal{S}$. Taking a union bound over the at most $s_{\max}$ true neighbours per node and over all p nodes, all true edges are retained with probability at least $1 - C_1 p s_{\max} \exp(-C_2 k\lambda^2)$, which tends to one under the stated scaling regime.

Bounding the number of selected edges. The skeleton is constructed from the thresholded support $\hat{\mathcal{S}}^{(\tau)} = \{i : |\hat{\beta}_i| > \tau\}$ with $\tau = \alpha\lambda$ for some $\alpha \in (0, 1)$, so that any variable with coefficient magnitude below the threshold is excluded. Since $\|\hat{\beta} - \beta^*\|_2 \geq \sqrt{\sum_{i\in\hat{\mathcal{S}}^{(\tau)}\setminus\mathcal{S}} |\hat{\beta}_i|^2} \geq \sqrt{|\hat{\mathcal{S}}^{(\tau)} \setminus \mathcal{S}|}\cdot\tau$, the thresholding condition $|\hat{\beta}_i| > \tau$ for $i \in \hat{\mathcal{S}}^{(\tau)}\setminus\mathcal{S}$ gives $\sqrt{|\hat{\mathcal{S}}^{(\tau)} \setminus \mathcal{S}|}\cdot\tau < \frac{\sqrt{s}}{\kappa}\lambda$, hence $|\hat{\mathcal{S}}^{(\tau)}| \leq (\frac{1}{\kappa\alpha} + 1)s$. Applying this bound to each of the p nodewise regressions and taking the union yields $|\hat{E}(\lambda)| \leq c_3\, p\, s_{\max}$ with high probability, for a constant c_3.

Combining the two parts, we obtain $\Pr(E^*_{\text{skel}} \subseteq \hat{E}(\lambda)) \geq 1 - C_1 p s_{\max} \exp(-C_2 k\lambda^2)$ and, on this event, $|\hat{E}(\lambda)| \leq c_3\, p\, s_{\max}$. □

C.7 Proof of Theorem 6.4

Proof of Theorem 6.4. Let $\mathcal{G}_{c_3}$ denote the class of DAGs whose skeleton is a subgraph of some edge set E with $E \supseteq E^*_{\text{skel}}$, $|E| \leq c_3 p s_{\max}$, and nodewise in-degree at most $s_{\max}$. By Theorem 6.3, on an event with probability tending to one,

$$E^*_{\text{skel}} \subseteq \hat{E}_{\text{skel}}, \qquad |\hat{E}_{\text{skel}}| \leq c_3 p s_{\max},$$

so $\mathcal{G}(\hat{E}_{\text{skel}}) \subseteq \mathcal{G}_{c_3}$ and $G^* \in \mathcal{G}(\hat{E}_{\text{skel}})$.

Fix any $G \neq G^*$ in $\mathcal{G}(\hat{E}_{\text{skel}})$ and define the set of changed nodes

$$\mathcal{J}(G, G^*) := \big\{ j : \ \mathrm{pa}_G(j) \neq \mathrm{pa}_{G^*}(j) \big\}.$$

Partition $\mathcal{J}$ into

$$\mathcal{J}_{\text{under}}(G, G^*) := \big\{ j \in \mathcal{J} : \mathrm{pa}_G(j) \not\supseteq \mathrm{pa}^*(j) \big\},$$
$$\mathcal{J}_{\text{over}}(G, G^*) := \big\{ j \in \mathcal{J} : \mathrm{pa}_G(j) \supsetneq \mathrm{pa}^*(j) \big\},$$

noting that $\mathcal{J} = \mathcal{J}_{\text{under}} \cup \mathcal{J}_{\text{over}}$ (disjoint). Write $A_j = \mathrm{pa}_G(j)$ and $A_j^* = \mathrm{pa}^*(j)$. The empirical score decomposes as

$$\widehat{\mathrm{Score}}_j^{\,\mathrm{dir}}(A) = \widehat{f}_j^{\mathrm{dir}}(A) + \mathrm{pen}(|A|),$$

where $\widehat{f}_j^{\mathrm{dir}}(A) = \frac{k}{2} \log \widehat{R}_{j|A}^{\mathrm{dir}}(u)$ is the empirical fit and $\mathrm{pen}(m) = \frac{1}{2}(\log k + 2\gamma_{\mathrm{EBIC}} \log p)\, m$ is the EBIC complexity penalty. The global score difference is

$$\widehat{S}^{\mathrm{dir}}(G) - \widehat{S}^{\mathrm{dir}}(G^*) = \sum_{j \in \mathcal{J}} \big[\widehat{f}_j^{\mathrm{dir}}(A_j) - \widehat{f}_j^{\mathrm{dir}}(A_j^*) + \mathrm{pen}(|A_j|) - \mathrm{pen}(|A_j^*|) \big].$$

By Lemma 6.1,

$$\eta_k := \max_{1 \le j \le p} \sup_A \big| \widehat{R}_{j|A}^{\mathrm{dir}}(u) - R_{j|A}^*(u) \big| = O_p\left(\sqrt{\frac{(s_{\max} + 1) \log p}{k}} \right) = o_p(1).$$

We show that each summand is strictly positive with probability tending to one.

Case I. Suppose $j \in \mathcal{J}_{\text{under}}$, so that $A_j \not\supseteq A_j^*$. By assumption (i), $R_{j|A_j}^*(u) - R_{j|A_j^*}^*(u) \geq \Delta_{\text{miss}}$. A mean-value expansion of $\widehat{f}_j^{\mathrm{dir}}(A) = \frac{k}{2} \log \widehat{R}_{j|A}^{\mathrm{dir}}(u)$ (possible since log is Lipschitz on $[r_{\min}(u)/2,\, 2r_{\max}(u)]$), gives

$$\widehat{f}_j^{\mathrm{dir}}(A_j) - \widehat{f}_j^{\mathrm{dir}}(A_j^*) = \frac{k}{2\, \bar{R}_j} \big[\widehat{R}_{j|A_j}^{\mathrm{dir}}(u) - \widehat{R}_{j|A_j^*}^{\mathrm{dir}}(u) \big] \geq \frac{k}{2\, \bar{R}_j} \big[\Delta_{\text{miss}} - 2\eta_k \big],$$

where $\bar{R}_j$ lies between $\widehat{R}^{\mathrm{dir}}_{j|A_j}(u)$ and $\widehat{R}^{\mathrm{dir}}_{j|A_j^*}(u)$. On the event $\eta_k < \min\{r_{\min}(u)/2,\, r_{\max}(u)\}$, all admissible empirical risks lie in $[r_{\min}(u)/2,\, 2r_{\max}(u)]$, hence $\bar{R}_j \le 2r_{\max}(u)$ and therefore

$$\widehat{f}_j^{\mathrm{dir}}(A_j) - \widehat{f}_j^{\mathrm{dir}}(A_j^*) \ge \frac{k}{4r_{\max}(u)}\big[\Delta_{\mathrm{miss}} - 2\eta_k\big].$$

The penalty difference satisfies

$$\big|\mathrm{pen}(|A_j|) - \mathrm{pen}(|A_j^*|)\big| \le \tfrac{1}{2}(\log k + 2\gamma_{\mathrm{EBIC}} \log p)\, s_{\max} = O(\log k \cdot s_{\max}).$$

Since $\frac{k\Delta_{\mathrm{miss}}}{4r_{\max}(u)} = \Omega(k)$ dominates $O(\log k \cdot s_{\max})$ for large k, the net score contribution is positive with probability tending to one.

Case II ($j \in \mathcal{J}_{\mathrm{over}}$): $A_j \supsetneq A_j^*$. By assumption (ii), $R^*_{j|A_j}(u) = R^*_{j|A_j^*}(u)$. Let $B_j := A_j \setminus A_j^*$, so $|B_j| \ge 1$. By Assumption 6.6, uniformly over such equal-risk supersets,

$$\widehat{f}_j^{\mathrm{dir}}(A_j^*) - \widehat{f}_j^{\mathrm{dir}}(A_j) = O_p(|B_j|).$$

The EBIC penalty gap is

$$\mathrm{pen}(|A_j|) - \mathrm{pen}(|A_j^*|) = \tfrac{1}{2}(\log k + 2\gamma_{\mathrm{EBIC}} \log p)\, |B_j|.$$

Hence

$$\widehat{\mathrm{Score}}_j^{\mathrm{dir}}(A_j) - \widehat{\mathrm{Score}}_j^{\mathrm{dir}}(A_j^*) \ge -O_p(|B_j|) + \tfrac{1}{2}(\log k + 2\gamma_{\mathrm{EBIC}} \log p)\, |B_j|.$$

Since $\log k + 2\gamma_{\mathrm{EBIC}} \log p \to \infty$, the penalty term dominates and the net score contribution is positive with probability tending to one.

Graph-level conclusion. Both cases yield strictly positive per-node contributions with probability tending to one. Hence

$$\widehat{S}^{\mathrm{dir}}(G) - \widehat{S}^{\mathrm{dir}}(G^*) > 0$$

for every $G \ne G^*$ with probability tending to one. Since $G^* \in \mathcal{G}(\hat{E}_{\mathrm{skel}})$, any direct-fit empirical minimizer satisfies $\widehat{G}^{\mathrm{dir}} = G^*$ with probability tending to one. □

C.8 Sensitivity Analysis to the Threshold

We assess how sensitive S3ME is to the choice of the exceedance threshold, which is parametrized as $q_{\text{conf}} = 1 - n^{\beta}/n$ for varying β. We fix $n = 1000$, $p = 50$, $m = 1$, and repeat 50 Monte Carlo replicates for each $\beta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, reporting mean skeleton F1/SHD and DAG F1/SHD. As β decreases, the threshold q_{conf} rises and the effective tail sample size k shrinks, making skeleton estimation harder, consistent with Theorem 6.3 which requires k to be large relative to $\log p$ for reliable sure screening. The orientation step exploits an asymptotic directional signal whose existence is guaranteed by Theorem 6.1 independently of k, and is therefore less sensitive to the threshold choice. Figure C.2 confirms both predictions. Skeleton F1 declines from 0.521 at $\beta = 0.9$ to 0.435 at $\beta = 0.5$, and skeleton SHD rises from 84.3 to 114.5 correspondingly, while DAG F1 remains stable at 0.869 and DAG SHD at 11.1 across $\beta \in \{0.6, 0.7, 0.8, 0.9\}$, with only mild degradation at $\beta = 0.5$ where k is smallest. This confirms that the orientation stage is robust to the threshold choice, and that the default $\beta = 0.7$ sits comfortably in the stable region.

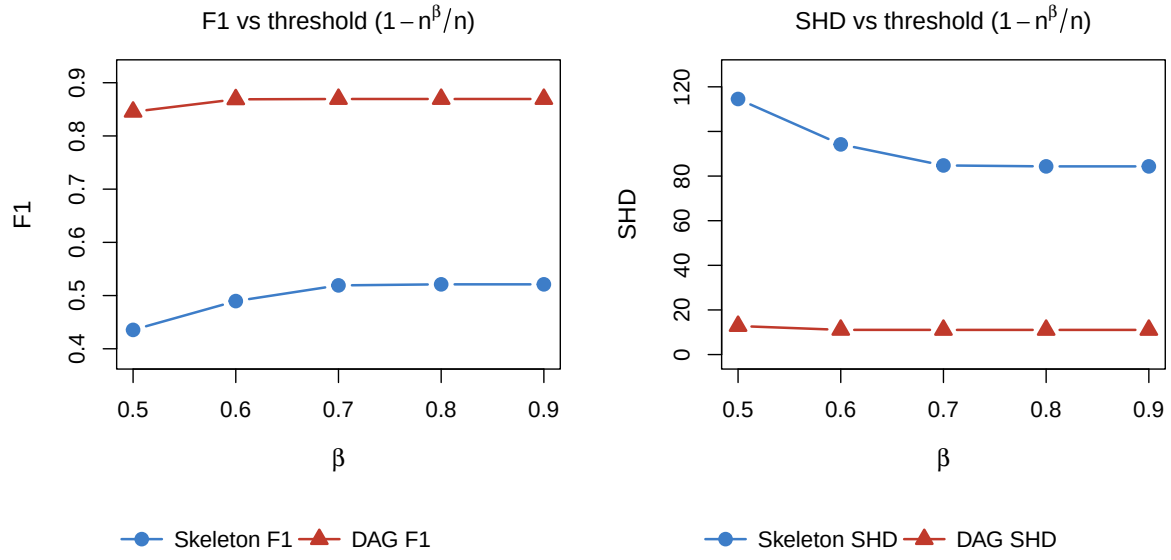

Figure C.2: Threshold sensitivity for $n = 1000$, $p = 50$, and 50 Monte Carlo replicates, where β sets the common exceedance quantile $q_{\text{conf}} = 1 - n^{\beta}/n$. The left panel shows mean $F_1 = 2PR/(P + R)$ and the right panel shows mean structural Hamming distance (SHD); blue and red denote the skeleton and full DAG stages. Here $P = \text{TP}/(\text{TP} + \text{FP})$, $R = \text{TP}/(\text{TP} + \text{FN})$, and the implemented SHD is $\text{FP} + \text{FN}$.

C.9 Robustness to Graph Structure

We examine whether the results from Section 5 depend on the choice of graph generating mechanism. All parameters are fixed at $n = 1000$, $p = 50$, conf_s = 10, $C = 1.0$, $\gamma_{\text{EBIC}} = 10$, and 50 Monte Carlo replicates. We compare Barabási–Albert (BA) graphs with Erdős–Rényi (ER) graphs (Erdős and Rényi, 1959), where the ER edge probability is set to $p_{\text{edge}} = 2m/(p-1)$ to align expected degree across graph types. Two attachment parameters are considered, $m = 1$ and $m = 2$.

Table C.1 reports skeleton and DAG metrics for three additional configurations, with the $m = 1$ BA results from Section 5 serving as the reference. DAG precision remains near 1.0 throughout, confirming that the orientation and pruning steps introduce very few spurious edges regardless of graph structure or attachment parameter. Performance differences are concentrated in recall. Under ER graphs, the uniform degree distribution allows the lasso to identify true neighbours more reliably, yielding DAG F1 of 0.979 and 0.887 for $m = 1$ and $m = 2$ respectively. The modest drop from $m = 1$ to $m = 2$ under ER reflects the expected difficulty of screening denser graphs. Under BA with $m = 2$, however, the same increase in density leads to a much sharper decline, with DAG F1 falling to 0.644 and DAG recall to 0.478. Since ER $m = 2$ and BA $m = 2$ have the same expected degree, this gap isolates the effect of the BA hub structure, where high-degree nodes create local neighbourhoods that are substantially harder to screen, causing true edges to be missed at the skeleton stage. The deficit then propagates to the DAG level. This confirms that the performance gap is attributable to graph topology rather than to a limitation of the method itself.

Table C.1: Graph-structure robustness for $n = 1000$, $p = 50$, and 50 Monte Carlo replicates, comparing Erdős–Rényi (ER) and Barabási–Albert (BA) graphs at attachment/density parameter m. Precision is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, recall is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, $F_1 = 2PR/(P+R)$, and the implemented SHD is $\mathrm{FP}+\mathrm{FN}$. Entries are Monte Carlo means with standard deviations in parentheses; the BA $m = 1$ baseline from Section 6.5 is omitted.

Setting	Stage	Precision	Recall	F1	SHD
ER, $m = 1$	Skeleton	0.212 (0.032)	0.963 (0.031)	0.346 (0.042)	182.82 (17.12)
	DAG	1.000 (0.000)	0.959 (0.032)	0.979 (0.017)	2.16 (1.89)
ER, $m = 2$	Skeleton	0.366 (0.045)	0.814 (0.062)	0.502 (0.040)	159.20 (20.86)
	DAG	1.000 (0.000)	0.800 (0.063)	0.887 (0.039)	20.12 (7.59)
BA, $m = 2$	Skeleton	0.217 (0.023)	0.522 (0.055)	0.306 (0.032)	229.72 (14.25)
	DAG	0.998 (0.007)	0.478 (0.056)	0.644 (0.053)	50.74 (5.59)

C.10 Application Preprocessing and Tuning Details

C.10.1 Danube preprocessing

For the Danube application, we use daily summer (June–August) discharge measurements from 31 gauging stations over the common period 1960–2010. To reduce short-range temporal dependence, we apply standard temporal declustering before marginal standardization. Each declustered margin is then transformed to unit Pareto scale by the empirical rank transform, and the log-tail coordinates are constructed as in Section 2.3 of the main paper. The application threshold is set at $q_{\mathrm{conf}} = 0.90$, which yields $k = 117$ tail observations.

C.10.2 S&P 500 preprocessing and tuning

For the S&P 500 application, the raw returns panel is obtained from Stooq over January 2005 to March 2025. Stocks with more than 5% missing observations are removed before analysis, leaving $p = 103$ constituents. Loss variables are defined by $L_{t,j} = -r_{t,j}$, after which each marginal series is transformed to unit Pareto scale by the empirical rank transform and the tail sample is selected using $q_{\mathrm{conf}} = 1 - n^{0.7}/n$, yielding $k = 4{,}120$ observations.

The skeleton is estimated by proxy-adjusted nodewise extremal LASSO with $\lambda = c\sqrt{\log(p+1)/k}$. Figure C.3 reports the total BIC score and resulting DAG sparsity across the tuning grid for the multiplier c. The elbow criterion selects $c = 5$, which yields the value used in the main-text application.

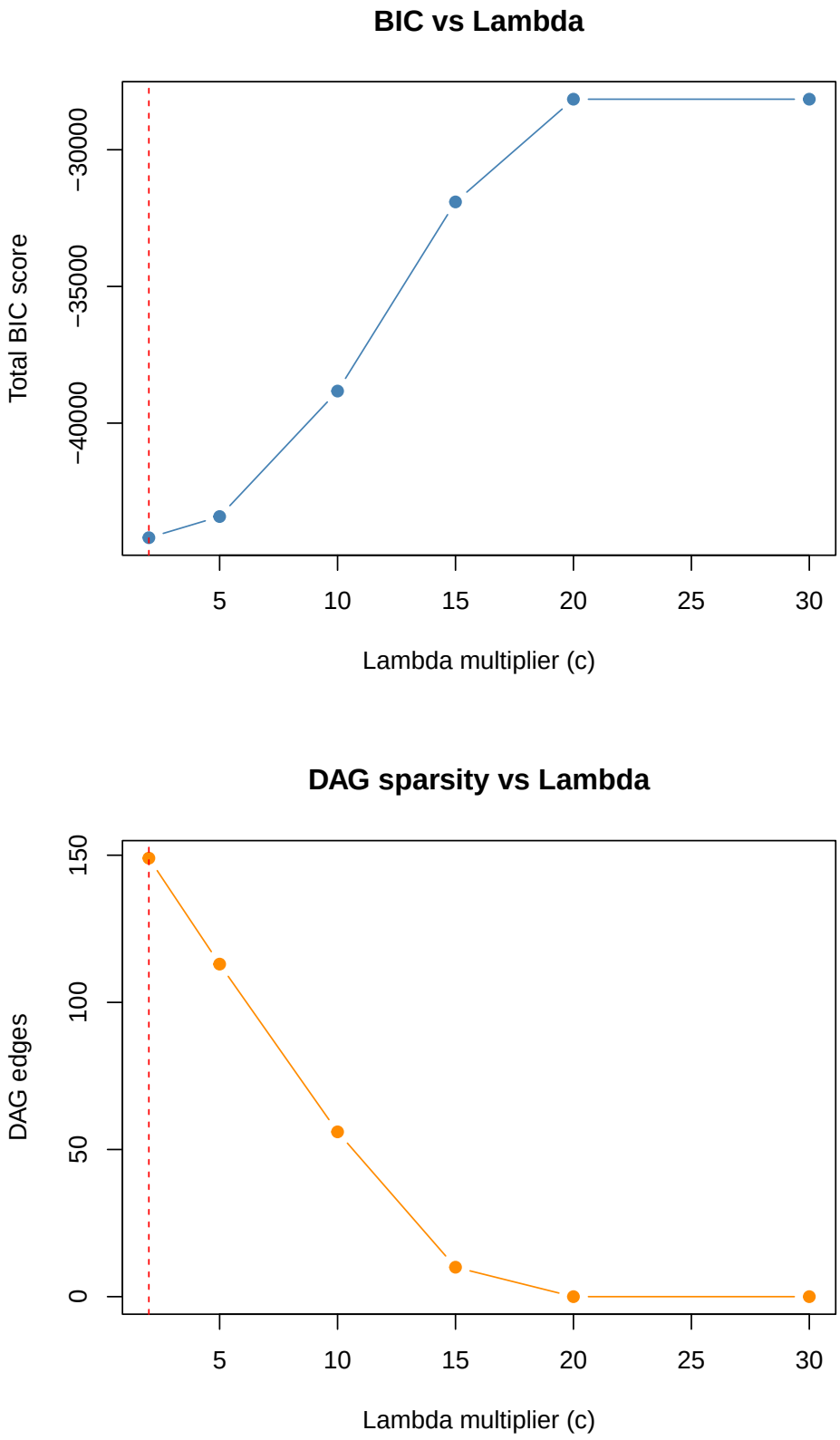

Figure C.3: Regularisation selection for the S&P 500 application. The upper panel shows the total Bayesian information criterion (BIC), for which smaller values indicate a preferred fit–complexity trade-off, against the multiplier c in $\lambda = c\sqrt{\log(p+1)/k}$. The lower panel shows the number of selected DAG edges. The elbow rule selects $c = 5$, marked by the red dashed line in both panels.

Bibliography

Allen, M. R. (2003). Liability for climate change. *Nature*, 421(6926):891–892.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton.

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *Ann. Appl. Stat.*, 9(4):2023–2050.

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley.

Bhuyan, P., Jana, K., and McCoy, E. J. (2023). Causal analysis at extreme quantiles with application to london traffic flow data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(5):1452–1474.

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Rashid, H., Uotila, P., Hirst, A., Kowalczyk, E., Golebiewski, M., Sullivan, A., et al. (2013). The ACCESS coupled model: description, control climate and evaluation. *Australian Meteorological and Oceanographic Journal*, 63(1):41–64.

Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I. I., Overland, J., Perlwitz, J., Sebbari, R., and Zhang, X. (2013). Detection and attribution of climate change: from global to regional. In Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen,

S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 867–952. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.

Bongiovanni, G., Matiu, M., Crespi, A., Napoli, A., Majone, B., and Zardi, D. (2025). EEAR-Clim: a high-density observational dataset of daily precipitation and air temperature for the Extended European Alpine Region. *Earth System Science Data*, 17(4):1367–1391.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Castro-Camilo, D. and Huser, R. (2020). Local likelihood estimation of complex tail dependence structures, applied to US precipitation extremes. *Journal of the American Statistical Association*, 115(531):1037–1054.

Cattiaux, J. and Ribes, A. (2018). Defining single extreme weather events in a climate perspective. *Bulletin of the American Meteorological Society*, 99(8):1557–1568.

Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. In Sims, C. A., editor, *Advances in Econometrics, Vol. 1: Sixth World Congress*, pages 171–209. Cambridge University Press, Cambridge.

Cheng, C. and Li, F. (2025). Inverting estimating equations for causal inference on quantiles. *Biometrika*, 112(1):asae058.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, 33:806–839.

Chernozhukov, V. and Fernández-Val, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies*, 78(2):559–589.

Chernozhukov, V., Fernández-Val, I., and Kaji, T. (2018). Extremal quantile regression: An overview. In Koenker, R., Chernozhukov, V., He, X., and Peng, L., editors, *Handbook of Quantile Regression*. Chapman and Hall/CRC.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Clarke, B., Otto, F. E. L., Stuart-Smith, R. R., and Harrington, L. J. (2022). Extreme weather impacts of climate change: an attribution perspective. *Environmental Research: Climate*, 1(1):012001.

Clemen, R. T., Reilly, T., and Winkler, R. L. (2000). Assessing dependence: Some experimental results. *Management Science*, 46(8):1100–1108.

Clemen, R. T. and Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research*, 33(2):427–442.

Clemen, R. T. and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London, London.

Coles, S. G., Heffernan, J., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3921–3962.

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.

Daouia, A., Gardes, L., and Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli*, 19(5B):2557–2589. Listed as Daouia et al. 2011 in the outline; this is the published version.

De Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer.

Deuber, D., Li, J., Engelke, S., and Maathuis, M. H. (2024). Estimation and inference of extremal quantile treatment effects for heavy-tailed distributions. *Journal of the American Statistical Association*, 119(547):2206–2216.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, pages 267–277.

Donald, S. G. and Hsu, Y.-C. (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics*, 178:383–397.

Engelke, S., Gnecco, N., and Röttger, F. (2025). Extremes of structural causal models. *arXiv preprint arXiv:2503.06536*.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932.

Engelke, S. and Taeb, A. (2025). Extremal graphical modeling with latent variables via convex optimization. *Journal of Machine Learning Research*, 26(42):1–68.

Engelke, S. and Volgushev, S. (2022). Structure learning for extremal tree models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(6):2055–2087.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958.

Fang, F., Bai, S., and Wang, T. (2025). Structural causal models for extremes: an approach based on exponent measures. *arXiv preprint arXiv:2508.00223*.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.

Fischer, E. M. and Knutti, R. (2015). Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. *Nature Climate Change*, 5(6):560–564.

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, volume 23, pages 604–612.

Frölich, M. and Melly, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357.

Gardes, L. and Girard, S. (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2):177–204.

Gillett, N. P. et al. (2016). The detection and attribution model intercomparison project (damip v1.0) contribution to cmip6. *Geoscientific Model Development*, 9:3685–3697.

Gissibl, N., Klüppelberg, C., and Lauritzen, S. (2021). Identifiability and estimation of recursive max-linear models. *Scandinavian Journal of Statistics*, 48(1):188–211.

Gissibl, N., Klüppelberg, C., and Otto, M. (2018). Tail dependence of recursive max-linear models with regularly varying noise variables. *Econometrics and Statistics*, 6:149–167.

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. (2021a). Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755–1778.

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. (2021b). Causal discovery in heavy-tailed models. *Ann. Stat.*, 49(3):1755–1778.

Gonzalez, P., Naveau, P., Thao, S., and Worms, J. (2025). A statistical method to model non-stationarity in precipitation records changes. *Geophysical Research Letters*, 52(6):e2023GL107201.

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

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464.

Hauser, M., Gudmundsson, L., Orth, R., Jézéquel, A., Haustein, K., Vautard, R., van Oldenborgh, G. J., Wilcox, L., and Seneviratne, S. I. (2017). Methods and model dependency of extreme event attribution: The 2015 european drought. *Earth's Future*, 5:1034–1043.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Hegerl, G. C., Hoegh-Guldberg, O., Casassa, G., Hoerling, M. P., Kovats, R. S., Parmesan, C., Pierce, D. W., and Stott, P. A. (2010). Good practice guidance paper on detection and attribution related to anthropogenic climate change. In Stocker, T. F., Field, C. B., Qin, D., Barros, V., Plattner, G.-K., Tignor, M., Midgley, P. M., and Ebi, K. L., editors, *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Detection and Attribution of Anthropogenic Climate Change*. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, pages 1163–1174.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 73–84. John Wiley & Sons, Ltd.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Holland, P. W. (1986). Statistics and causal inference: Rejoinder. *Journal of the American Statistical Association*, 81(396):968. DOI: 10.2307/2289069 MAG ID: 2326415568.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jézéquel, A., Dépoues, V., Guillemot, H., Trolliet, M., Vanderlinden, J.-P., and Yiou, P. (2018). Behind the veil of extreme event attribution. *Climatic Change*, 149(3):367–383.

Jiang, J., Richards, J., Huser, R., and Bolin, D. (2025). Separation-based causal discovery for extremes. *arXiv preprint arXiv:2505.08008*.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245.

Kirchmeier-Young, M. C., Wan, H., Zhang, X., and Seneviratne, S. I. (2019). Importance of framing for extreme event attribution: The role of spatial and temporal scales. *Earth's Future*, 7:1192–1204.

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

Kiriliouk, A., Rootzén, H., Segers, J., and Wadsworth, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61(1):123–135.

Klüppelberg, C. and Lauritzen, S. L. (2019). Bayesian networks for max-linear models. In Biagini, F., Kauermann, G., and Meyer-Brandis, T., editors, *Network Science: An Aerial View*, pages 79–97. Springer International Publishing, Cham.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, Cambridge.

Koenker, R. W. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Krali, M. (2025). Causal discovery in heavy-tailed linear structural equation models via scalings. *Scandinavian Journal of Statistics*.

Lanet, M., Li, L., Ehret, A., et al. (2024). Attribution of summer 2022 extreme wildfire season in southwest france to anthropogenic climate change. *npj Climate and Atmospheric Science*, 7:267.

Lauritzen, S. L. (1996). *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. Oxford University Press, Oxford.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):475–499.

Lehmann, E. L. and D'Abrera, H. J. (1975). *Nonparametrics: statistical methods based on ranks*. Toronto: Holden-Day.

Li, M. and Castro-Camilo, D. (2025). *eFCM*. R package version 1.0.

Li, M. et al. (2025). Extreme event attribution via causal inference: Methods and applications. *Journal of the Royal Statistical Society Series B*. In preparation.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15:3065–3105.

Ma, H., Qin, J., and Zhou, Y. (2023). From conditional quantile regression to marginal quantile estimation with applications to missing data and causal inference. *Journal of Business & Economic Statistics*, 41(4):1377–1390.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., et al. (2021). Climate Change 2021: The Physical Science Basis. Contribution of working group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *IPCC: Geneva, Switzerland*.

McPhillips, L. E., Chang, H., Chester, M. V., Depietri, Y., Friedman, E., Grimm, N. B., Kominoski, J. S., McPhearson, T., Méndez-Lázaro, P., Rosi, E. J., et al. (2018). Defining extreme events: A cross-disciplinary review. *Earth's Future*, 6(3):441–455.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, 34(3):1436–1462.

Mhalla, L., Chavez-Demoulin, V., and Dupuis, D. J. (2020). Causal mechanism of extreme river discharges in the upper danube basin network. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(4):741–764.

Montiel, L. V. and Bickel, J. E. (2012). A simulation-based approach to decision making with partial information. *Decision Analysis*, 9(4):329–347.

Nanditha, J., Villarini, G., Kim, H., and Naveau, P. (2024). Strong linkage between observed daily precipitation extremes and anthropogenic emissions across the contiguous United States. *Geophysical Research Letters*, 51(20):e2024GL109553.

National Academies of Sciences, Engineering, and Medicine (2016). *Attribution of Extreme Weather Events in the Context of Climate Change*. The National Academies Press, Washington, DC.

Naveau, P., Hannart, A., and Ribes, A. (2020a). Statistical methods for extreme event attribution in climate science. *Annual Review of Statistics and Its Application*, 7:89–110.

Naveau, P., Hannart, A., and Ribes, A. (2020b). Statistical methods for extreme event attribution in climate science. *Annual Review of Statistics and Its Application*, 7(1):89–110.

Neyman, J. (1990). On the application of probability theory to agricultural experiments: Essay on principles. Section 9. *Statistical Science*, 5(4):465–472. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in *Roczniki Nauk Rolniczyc*, Tom X (1923): 1–51.

Noy, I. et al. (2023). Event attribution is ready to inform loss and damage negotiations. *Nature Climate Change*, 13:1279–1281.

Otto, F. E. L. (2017). Attribution of weather and climate events. *Annual Review of Environment and Resources*, 42:627–646.

Otto, F. E. L., Massey, N., van Oldenborgh, G. J., Jones, R. G., and Allen, M. R. (2012). Reconciling two approaches to attribution of the 2010 Russian heat wave. *Geophysical Research Letters*, 39:L04702.

Paciorek, C. J., Stone, D. A., and Wehner, M. F. (2018). Quantifying statistical uncertainty in the attribution of human influence on severe weather. *Weather and climate extremes*, 20:69–80.

Pasche, O. C., Chavez-Demoulin, V., and Davison, A. C. (2023). Causal modelling of heavy-tailed variables and confounders with application to river flow. *Extremes*, 26(3):573–594.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96 – 146.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition.

Peng, L. and Fine, J. P. (2009). Competing risks quantile regression. *Journal of the American Statistical Association*, 104(488):1440–1453.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

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

Philip, S. Y., Kew, S. F., van Oldenborgh, G. J., Otto, F. E. L., Vautard, R., van der Wiel, K., King, A. D., Lott, F. C., Arrighi, J., Singh, R., and van Aalst, M. K. (2020). A protocol for probabilistic extreme event attribution analyses. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(2):177–203.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*, volume 4 of *Applied Probability*. Springer-Verlag, New York.

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer.

Risser, M. D. and Wehner, M. F. (2017). Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey. *Geophysical Research Letters*, 44:12457–12464.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Multivariate generalized pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131.

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

Rosenbaum, P. R. and Rubin, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Runge, J., Bathiany, S., Bollt, E. M., Camps-Valls, G., Coumou, D., Deyle, E. R., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):2553.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, 1 edition.

Shepherd, T. G. (2016). A common framework for approaches to extreme event attribution. *Current Climate Change Reports*, 2(1):28–38.

Shepherd, T. G., Boyd, E., Calel, R. A., Chapman, S. C., Dessai, S., Dima-West, I. M., Fowler, H. J., James, R., Maraun, D., Martius, O., Senior, C. A., Sobel, A. H., Stainforth, D. A., Tett, S. F. B., Trenberth, K. E., van den Hurk, B. J. J. M., Watkins, N. W., Wilby, R. L., and Zenghelis, D. A. (2018). Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. *Climatic Change*, 151(3):555–571.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E., and Knutti, R. (2020). Climate change now detectable from any single day of weather at global scale. *Nature Climate Change*, 10(1):35–41.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition.

Stott, P. A., Stone, D. A., and Allen, M. R. (2004). Human contribution to the European heatwave of 2003. *Nature*, 432:610–614.

Tran, N. M., Buck, J., and Klüppelberg, C. (2024). Estimating a directed tree for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 86(3):771–792.

Trenberth, K. E., Fasullo, J. T., and Shepherd, T. G. (2015). Attribution of climate extreme events. *Nature Climate Change*, 5(8):725–730.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

Van Der Vaart, A. W. and Wellner, J. A. (2023). Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer.

Van Der Wiel, K., Kapnick, S. B., Van Oldenborgh, G. J., Whan, K., Philip, S., Vecchi, G. A., Singh, R. K., Arrighi, J., and Cullen, H. (2017). Rapid attribution of the August 2016 flood-inducing extreme precipitation in south Louisiana to climate change. *Hydrology and Earth System Sciences*, 21(2):897–921.

van Oldenborgh, G. J., van der Wiel, K., Kew, S., Philip, S., Otto, F., Vautard, R., King, A., Lott, F., Arrighi, J., Singh, R., et al. (2021). Pathways and pitfalls in extreme event attribution. *Climatic Change*, 166(1):1–27.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., et al. (2019). Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, 11(7):2177–2213.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Wang, H. J., Li, D., and He, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, 107(500):1453–1464.

World Economic Forum (2022). The Global Risks Report 2022 (17th ed.). World Economic Forum.

Xu, W., Wang, H. J., and Li, D. (2022). Extreme quantile estimation based on the tail single-index model. *Statistica Sinica*, 32(2):893–914.

Zhang, Y. (2018a). Extremal quantile treatment effects. *The Annals of Statistics*, 46(6B):3707–3740.

Zhang, Y. (2018b). Supplement to "extremal quantile treatment effects.".

Zhang, Z., Chen, Z., Troendle, J. F., and Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics*, 68(3):697–706.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., Bretonière, P.-M., Brönimann, S., Casanueva, A., Cherchi, A., Davin, E. L., Hegerl, G. C., Jakob, C., Katragkou, E., Kim, W. C., Kröner, N., Lemonnier, F., Messori, G., Muñoz, Á. G., O'Gorman, P. A., Rohmer, J., Rösch, M., Sippel, S., Shkolnik, I., Szekely, E., Vignotto, E., Wehner, M., and Zlobina, D. (2020). A typology of compound weather and climate events. *Nature Reviews Earth & Environment*, 1:333–347.