Extreme value theory and graphs, bridging the gap

Aiden Farrell B.Sc.(Hons), M.Sc.

Submitted for the degree of Doctor of Philosophy at
Lancaster University

September 2025

Abstract

Over the last twenty years, significant advancements have been made in extreme value theory (EVT), many of which have been achieved by leveraging concepts from other mathematical areas. One such example in the last five years is graph theory. This has gone both ways: extreme value theory has been used to model characteristics of graphical structure, such as their degree distribution, and graphs have been used for their stochastic representation to aid in dimensionality reduction in EVT models. The work contained in this thesis contributes to both of these areas.

For characteristics of graphical structures, modelling the degree distribution (the number of edges connected to a vertex) will be our aim. In the literature, such modelling has several limitations. Some authors approximate the degree distribution, a discrete random variable, with a continuous distribution, while others use a single discrete model for the entire degree distribution, despite the data exhibiting different behaviour in the body and the tail. More concerning is the restriction of most analyses to heavy-tailed distributions. Given these issues, we propose a flexible mixture distribution, where the tail is modelled using an integer generalised Pareto distribution, to model the *entire* degree distribution. Consequently, the model can capture a variety of behaviour in both the body and the tail, without needing to pre-specify the rate of tail decay.

For exploiting the stochastic representation of graphs, this has been achieved under the assumption of full asymptotic dependence (AD). For data exhibiting asymptotic independence (AI), the theory has been developed, but there is no statistical methodology to accompany it. We aim to develop such methodology. We achieve this in a different way than the theory suggests for computational purposes, by proposing an extension to the conditional multivariate extreme value model (CMEVM). Our extension has three ingredients: a new model for the margins of the residual distribution, a novel approach for incorporating graphical structures into the dependence structure of the residual distribution, and a step-wise inference procedure that loses no information compared to a joint estimation procedure to allow for

scalable inference to high dimensions. Our results indicate the necessity for a general graphical dependence structure and a flexible dependence model when applied to river discharges in the upper Danube River basin.

Treating river discharges as a multivariate problem meant we were unable to obtain predicted river flows at unobserved locations on the river network without post-hoc interpolation, which may not accurately represent river flow, particularly at confluence points. This inspired modelling river discharges as a stochastic process on a non-Euclidean space. However, this has received little attention in the EVT literature, and where it has, the model has assumed full AD. Thus, we propose a further extension of the CMEVM that models the residual distribution using the very recently developed class of Gaussian Whittle-Matérn fields for metric graphs. By treating the data in this manner, we can form a geodesic distance metric on the graph and use a correlation function based on river distance to capture the dependence between locations. Consequently, we can obtain fast and fully stochastic simulations for any point on the river network. Although our results are preliminary, they offer valuable insight into potential advancements in stochastic simulations of river flows.

Acknowledgements

I would like to start by acknowledging my academic supervisors, Emma Eastoe and Clement Lee. Throughout my PhD journey, you have both provided unwavering support, a tremendous amount of patience and understanding, and insightful counsel. You have both pushed me to learn more about statistics and extreme value theory than I thought possible, and your ability to break these difficult concepts down into simple constructs has helped me beyond words. I could not imagine completing my thesis with anyone else, and this thesis would not be nearly as complete without your guidance and helpful comments. Emma, I would like to personally thank you for introducing me to the world of extreme value theory, guiding me through my Master's project, and welcoming me back into research. Clement, I would like to thank you for expanding my horizons with the applications of extremes, which ultimately shaped this entire thesis. Returning to Lancaster to work with you both has been hugely rewarding and is an opportunity that I will be forever grateful for.

Lancaster University has a large community dedicated to researching extreme value theory, and I have been privileged enough to join this community and to have been invited to join the Extremes Reading Group. Thank you to everyone who has participated in the group during my time for providing thought-provoking seminars and insightful discussions that improved my own work. While we were a reading group, the social elements, ranging from hikes in Lake Como, meals and coffee in Lancaster, and attending baseball games in North Carolina, provided us with a reprieve from research and allowed us to develop friendships that I hope last into the future.

Speaking of reprieves from research, I would like to thank Alin, Jess, Cían, and Charlotte for adopting me into their Bayesian statistics group and providing much-needed breaks with our numerous coffees at Café Republic and several wine nights. To Alex and Rachael, I am very grateful to have met you both during our Master's year. Without you, I would never have completed that course, let alone the PhD, so thank you. During my PhD, I was elected as President of Graduate College Bar Sports. While President, I have gained numerous friends,

whom I will cherish forever. To my team, “Up the Owl!” and to the wider Bar Sports community, thank you for cultivating my love of pool, darts, and dominoes. Thank you also to my quiz team “And In Last Place” for teaching me things outside of statistics, and not chiding me when I got maths-based questions incorrect. Although generally not providing a reprieve from research, as our conversations would always converge to mathematics in some form or another, I would like to thank Andrew for supporting me, both before and during the PhD. Conversations with you were always helpful, whether you were giving me perspective or aiding my plans for the future, and I am excited for our friendship to continue in London.

Last, but certainly not least, I would like to thank my family. My decision to leave London and undertake a PhD at Lancaster may have seemed abrupt and strange, but you have always supported me in my journey, as long as I was happy. Thank you for listening to me, reassuring me that everything will be okay, and reminding me to take a break when needed. Without you, I would not be where I am today, and for that, I cannot thank you enough.

Declaration

I declare that the work in this thesis is my own, unless otherwise stated, and has not been submitted elsewhere for the award of any other degree.

Chapters 2, 3, and 4 are joint work with Dr. Emma F. Eastoe and Dr. Clement Lee.

Chapter 2, although unpublished, is adjacent to the paper of Lee. C, Eastoe F. E., and Farrell, A. (2024). Degree distributions in networks: Beyond the power law. Statistica Neerlandica. https://doi.org/10.1111/stan.12355.

Chapter 3 is a draft paper to be submitted for publication as Farrell, A., Eastoe, E. F., and Lee, C. (2025). Conditional Extremes with Graphical Models.

Chapter 5 is a result of the Lancaster and Maynooth University contribution for a competition as part of the 2023 Extreme Value Analysis conference at the University of Bocconi, Italy. This has been published as André, L. M., Campbell, R., D'Arcy, E., Farrell, A., Healy, D., Kakampakou, L., Murphy, C., Murphy-Barltrop, C. J. R., and Speers, M. (2024). Extreme value methods for estimating rare events in Utopia. Extremes. https://doi.org/10.1007/s10687-024-00498-w. My contributions are in Sections 5.1 - 5.3.

I would like to acknowledge comments from various anonymous referees, which have aided in the improvement of the work in these chapters.

The word count for this thesis is approximately $43,486$.

Aiden Farrell

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

This chapter outlines the core theme of the thesis: to combine extreme value theory (EVT) and graph theory. Since the motivation for each chapter is slightly different, some motivation is provided in the thesis outline below, with a more thorough description provided in the introduction of the requisite chapters. In addition, brief literature reviews of both EVT (univariate, multivariate, and spatial) and graph theory are provided, with more comprehensive reviews and introductions into the necessary methods presented in each chapter.

1.1 Thesis outline

This thesis aims to combine two areas of mathematics: graph theory and EVT. Given the distinct nature of Chapters 2 - 4, which focus on univariate, multivariate, and spatial problems, respectively, the chapters themselves are treated as self-contained topics. Consequently, the notation between chapters may not be consistent, however, the notation within each chapter should be consistent and well-defined within said chapter. Furthermore, individual discussion sections and reference lists are provided at the end of each chapter.

Chapter 2 aims to model the degree distribution of a network using methods from EVT. Previous attempts to model the degree distribution have various limitations. Firstly, existing methods to model the *entire* degree distribution are generally not flexible enough to capture the differing behaviour in the body and the tail. Secondly, most existing models assume a heavy-tailed distribution for the tail, which is not appropriate for all networks. Further, there is a limited pool of generative models that can produce a network with a heavy-tailed degree distribution. To overcome these limitations, we draw on EVT to model degree distributions using a flexible mixture model. This allows us to model the *entire* degree distribution, account for the differing behaviours of the body and the tail, and have a flexible model that can capture light-, exponential-, and heavy-tails without having to pre-specify the rate of tail decay. We illustrate our proposed method on a range of degree distributions to explore its merits and limitations.

Chapters 3 and 4 switch focus. Rather than using extreme value methods to model data from networks, we use graphs to improve extreme value models. Chapter 3 focuses on the multivariate setting, using conditional independence implied from graphs to reduce the number of dependence parameters in multivariate tail models. This work was inspired by Engelke and Hitz (2020), who developed extremal graphical models for multivariate Pareto distributions. Since this distribution can only capture full asymptotic dependence (AD), Casey and Papastathopoulos (2023) extended the concept to the conditional multivariate extreme value model Heffernan and Tawn (2004), which can capture both AD and asymptotic independence

(AI). However, Casey and Papastathopoulos (2023) do not go as far as to use this theory for statistical inference. We fill this gap by proposing several extensions to the CMEVM, including a more flexible model for the margins of the residual distribution, a sparse dependence structure for the residual distribution, and a step-wise inference procedure to allow for scalable inference to high dimensions without loss of information. We apply our proposed model to river discharges in the upper Danube River basin to demonstrate the advantages of a more general and flexible model for the dependence structure.

In Chapter 3, the flows at different measurement locations in the upper Danube River basin are modelled as a multivariate random vector, meaning predictions of river flow at unobserved locations on the river can only be obtained with post-inference interpolation. This limitation suggests the need for a process-based model. While there exist many extreme value models for stochastic processes observed on Euclidean spaces, there are very few that can model stochastic processes observed on non-Euclidean spaces. One example is the model proposed by Asadi et al. (2015) that uses a Brown-Resnick process in which both Euclidean and non-Euclidean between-site distances contribute to the pairwise dependence between locations. However, this model assumes full AD, which we show in Chapter 3 to be inappropriate for this dataset. In addition, although their model fits relatively well, extrapolation to unobserved locations requires considerable data processing using a digital elevation model. This inspires our next extension of the CMEVM, which uses generalised additive models (GAMs) to describe the marginal and CMEVM parameters, and a Gaussian Whittle-Matérn field for metric graphs (Bolin et al., 2024) to model the residual process. Our model allows for fast predictions across the entire river network.

Finally, Chapter 5 details Lancaster University's contribution to the EVA 2023 Conference Data Challenge. The challenge consisted of several univariate (C1 and C2) and multivariate (C3 and C4) problems. For the former, flexible GAMs for extreme value distributions are used to model a non-stationary time series to estimate extreme quantiles. For the latter, two methods were used due to the differing datasets. For C3, we propose modelling a GAM extension of the approach proposed by Wadsworth and Tawn (2013) due to the required AI assumption. For C4, a clustering technique, based on exploratory analysis of pairwise extremal dependence measures, is used to reduce a 50-dimensional random vector into 5 distinct vectors of at most dimension 15. The clusters are then modelled using the original CMEVM and are subsequently used to estimate high-dimensional probabilities while assuming independence between the clusters.

1.2 Univariate extreme value theory

Extreme value theory aims to characterise the tail behaviour of any probability distribution without knowing the form of the underlying distribution. This allows us to develop statistical methodology for modelling the tails of the distribution without using information in the body. In the univariate case, the two main approaches for modelling extreme values are: the generalised extreme value (GEV) distribution, described in Section 1.2.1 for modelling block maxima, and the generalised Pareto distribution (GPD), described in Section 1.2.2 for modelling threshold exceedances. A review of these models can be found in Coles (2001). Both approaches assume a sequence of independent and identically distributed (IID) random variables $X_1, \ldots, X_n$ for $n \in \mathbb{N}$ with common *continuous* distribution function F.

1.2.1 Generalised extreme value distribution

For the block maxima approach, we are concerned with the behaviour of the random variable $M_n = \max\{X_1, \ldots, X_n\}$, the maximum of the sequence. Minima can be modelled similarly by noting that $\min\{X_1, \ldots, X_n\} = -\max\{-X_1, \ldots, -X_n\}$. The distribution for M_n can be derived exactly for all $n \in \mathbb{N}$ and all $x \in \mathbb{R}$ as

$$\mathbb{P}[M_n \leq x] = \mathbb{P}[X_1 \leq x, \ldots X_n \leq x] = \mathbb{P}[X_1 \leq x] \ldots \mathbb{P}[X_n \leq x] = \{F(x)\}^n. \tag{1.2.1}$$

Equation (1.2.1) is not useful for constructing statistical models since F is unknown in practice and the distribution of M_n is degenerate to a point mass on x^F, the upper end-point of the distribution F since $\{F(x)\}^n \to 0$ as $n \to \infty$ for any $x < x^F$. To overcome this, consider the Extremal Types Theorem (Leadbetter, 1983). This states that if $X_1, \ldots X_d$ are IID random variables with common distribution function F, and there exists sequences of normalising constants $\{a_n\}$ and $\{b_n > 0\}$ such that for all $x \in \mathbb{R}$

$$\mathbb{P}\left[\frac{M_n - a_n}{b_n} \leq x\right] = \{F(a_n x + b_n)\}^n \to G(x) \quad \text{as } n \to \infty, \tag{1.2.2}$$

where G is a non-degenerate distribution function, then G takes the form of the GEV distribution with distribution function

$$G(x) = \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} & \text{if } \xi \neq 0 \\ \exp\left\{-\exp\left[-\frac{(x-\mu)}{\sigma}\right]\right\} & \text{if } \xi = 0 \end{cases} \tag{1.2.3}$$

such that $\sigma > 0$ and $A_+ = \max\{0, A\}$.

The parameters of the GEV distribution are $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$ corresponding to the location, scale, and shape, respectively. The case when $\xi = 0$ is taken in the limit as $\xi \to 0$. The common distribution function F is said to lie in the *domain of attraction* (DOA) of G, and the value of ξ is determined by the rate of decay of F. If $\xi < 0$, F has a finite upper end-point (light tail) and belongs to the DOA of the negative Weibull distribution. If $\xi > 0$, F has no upper endpoint (heavy tail) and belongs to the DOA of the Fréchet distribution. Finally, in the limit as $\xi \to 0$, F has no upper endpoint (exponential tail) and belongs to the DOA of the Gumbel distribution.

The normalising functions $\{a_n\}$ and $\{b_n > 0\}$ depend on F, which is unknown, however, they do not need to be modelled. Assuming equation (1.2.3) holds for all $x \in \mathbb{R}$ and large n, we can write

$$\mathbb{P}[M_n \leq x] \approx G\left(\frac{x - a_n}{b_n}\right) = G^*(x),$$

where G^* is a member of the GEV family with parameters μ^*, σ^* and ξ, then the normalising constants are absorbed into the location and scale parameters. Therefore, we can use the GEV distribution to model not only maxima of a sequence, but block maxima. To obtain the block maxima dataset, we split the data into m blocks of size n. Typically for environmental data, the blocks are taken to correspond to a time length of one year, such that n is the number of observations in a year and the block maxima are annual maxima. Taking the maximum in each block leads to the sequence $M_{n,1}, \ldots, M_{n,m}$ which are assumed to follow a GEV distribution. The approach is illustrated in Figure 1.1 (left panel).

Note, the choice of n (m) involves a bias-variance trade-off; if n (m) is too small (large), there are not enough points within each block such that the limiting result in equation (1.2.2) holds. Conversely, if n (m) is too large (small), there are not enough block maxima to accurately estimate the model's parameters.

1.2.2 Generalised Pareto distribution

The block maxima approach can lead to many, possibly informative, extreme observations being discarded during inference. This is common when several extreme values may occur in the same block, particularly if the block length is large, or when we model data at an aggregated level, i.e. we may only model the daily maxima, but the data may be recorded hourly. To overcome this limitation, an alternative is the peaks over threshold approach that treats all events above a large threshold as extreme. This approach is illustrated in Figure 1.1 (right panel).

Specifically, consider X to be an arbitrary element of the sequence of IID random variables

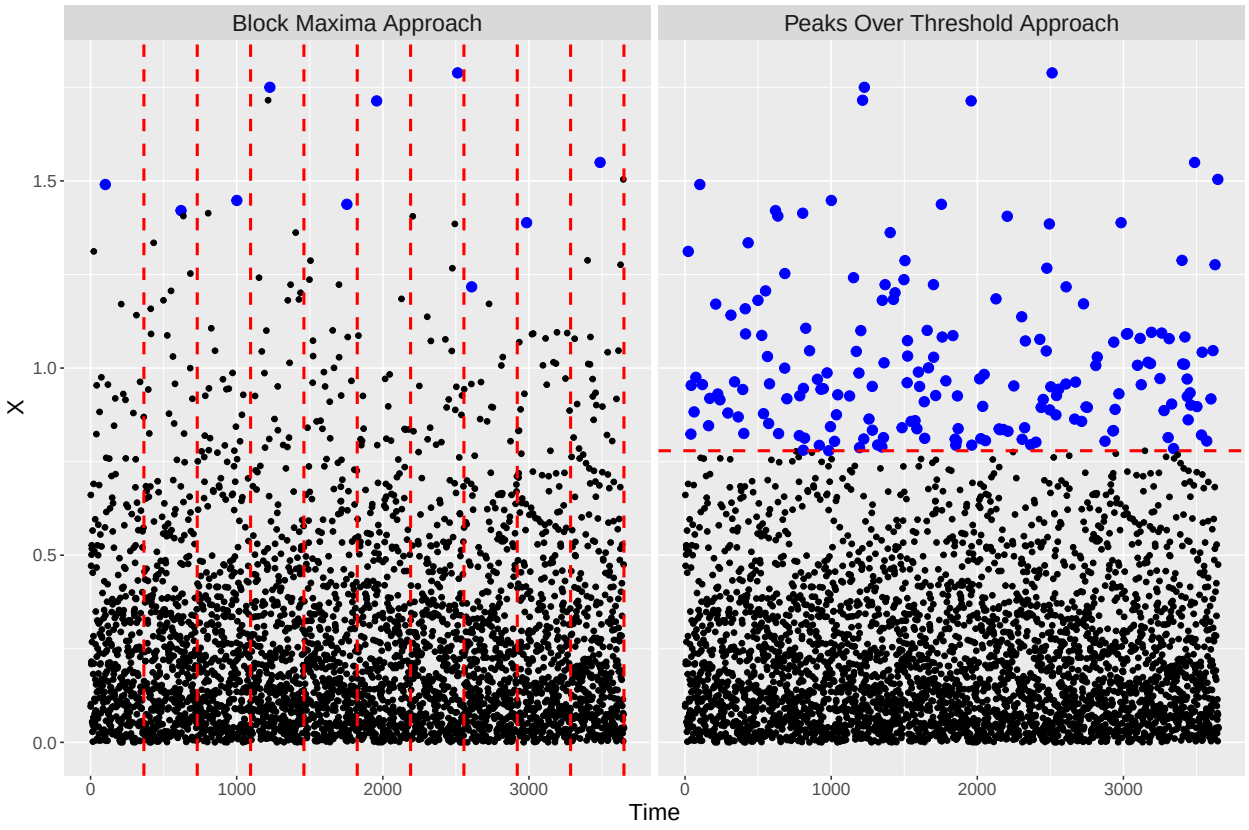

Figure 1.1: Block maxima (left) and peaks over threshold (right) approaches. The blue points represent the observations used in each modelling approach. The red dotted lines represent the block boundaries for the block maxima approach and the threshold for the peaks over threshold approach.

$X_1, \ldots, X_n$ with common *continuous* distribution function F and upper end-point x^F. If F is in the DOA of the GEV, the Pickands-Balkema-de Haan theorem (Balkema and de Haan, 1974; Pickands III, 1975) states that, under certain regularity conditions, there exists a normalising function $c(u) > 0$ such that for all $x > 0$

$$\mathbb{P}\left(\frac{X-u}{c(u)} \leq x \mid X > u\right) \to H(x) \tag{1.2.4}$$

as $u \to x^F$. Here, H is the distribution function of the generalised Pareto Distribution (GP) that takes the form

$$H(x) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp\left[-(x-u)/\sigma_u\right] & \text{if } \xi = 0 \end{cases} \tag{1.2.5}$$

where $\sigma_u > 0$ and $\xi \in \mathbb{R}$ are the scale and shape parameters, respectively. Again, $\xi = 0$ in equation (1.2.5) is taken to be in the limit. Note that the shape parameter ξ is the same as

the GEV shape parameter in equation (1.2.3), however, the scale parameter is not consistent as $\sigma_u = \sigma + \xi(u - \mu)$ and therefore depends on the location and scale parameter of the GEV, and the threshold u.

As with the sequence of normalising constants in the block maxima approach, the normalising function $c(u)$ depends on F and is unknown in practice. Assuming limit 1.2.4 holds for some large threshold u, the excess above the threshold can be modelled as

$$\mathbb{P}[X - u \leq x \mid X > u] \approx H\left(\frac{x}{c(u)}\right) = H^*(x),$$

where H^* is the distribution function of another GP distribution with scale parameter σ_u^* and shape parameter ξ. Again, the normalising function is absorbed into the scale parameter, allowing us to use equation (1.2.5) to model excesses above u. Smith (1989) explains the relationship between the GEV and GP distributions, and Coles (2001) provides a formal justification for this model in describing threshold exceedances.

Similar to the choice of n (m) in the block maxima approach, the choice of u represents another bias-variance trade-off (Smith, 1987). If u is chosen to be too low, limit (1.2.4) will not hold, resulting in biased parameter estimates, while too high a threshold will result in high variability of the estimates. There are many methods to choose the threshold; see Scarrott and MacDonald (2012) for a review. One common method detailed in Coles (2001) is to select the threshold based on mean residual life plots and parameter stability plots (Davison and Smith, 1990), with the former being linear and the latter being constant above a high threshold u. These methods are subjective, and even experienced practitioners can disagree on the threshold. Therefore, numerous other automated threshold selection procedures have been proposed, including likelihood-based diagnostic tools (Wadsworth, 2016), Bayesian cross-validation (Northrop et al., 2016), and Monte-Carlo-based approaches (Murphy et al., 2025; Varty et al., 2021).

1.3 Multivariate extreme value theory

Multivariate extreme value theory initially mirrored the univariate setting with models built for componentwise maxima and then for peaks over threshold data. However, both methods are limited in the types of joint extreme behaviour they can capture. To explain why and how it can be addressed, we first define the different classes of extremal dependence before presenting the conditional multivariate extreme value model.

1.3.1 Dependence measures

A key part of multivariate extreme value models is accurately capturing the extremal dependence structure. Let $V = \{1, \ldots, d\}$ and consider the d-dimensional random vector $X = \{X_j : j \in V\}$ with joint and marginal distributions $F_X(x) = \mathbb{P}[X \leq x]$ and $F_j(x_j) = \mathbb{P}[X_j \leq x_j]$, respectively. Simpson et al. (2020) consider a multivariate version of the coefficient χ (Joe, 1997) to determine if the extreme events of two or more components of X can occur simultaneously or not. Now, for $u \in (0,1)$, $A \subseteq V$ and $|A| \geq 2$, if the measure

$$\chi_A = \lim_{u \to 1} \chi_A(u) = \lim_{u \to 1} \mathbb{P}[F_i(X_i) > u : i \in A]/(1-u), \tag{1.3.1}$$

is strictly positive, then the components in A experience their extremes simultaneously and are said to belong to the extremal dependence class of asymptotic dependence (AD). However, if $\chi_A = 0$, then the components of X cannot be simultaneously extreme. Furthermore, if $|A| = 2$ the components are said to belong to the extremal dependence class of asymptotic independence (AI) (Ledford and Tawn, 1996). Full AD occurs when $\chi_A > 0$ for all subsets $A \subseteq V$, and full AI occurs when $\chi_A = 0$ for all two-dimensional subsets of V. While independence implies AI, the converse is not true.

When $|A| = 2$, χ_A does not tell us about the strength of the association between the components when the components exhibit AI. $i, j \in A$, $i \neq j$, and $u \in (0,1)$ the measure

$$\bar{\chi} = \lim_{u \to 1} \bar{\chi}(u) = \lim_{u \to 1} 2\log(1-u)/\mathbb{P}[(F_i(X_i) > u, F_j(X_j) > u)],$$

quantifies the strength of association between the components (Coles et al., 1999). Since our focus is random variables with AI, we would need to analyse $(\chi, \bar{\chi})$ together to determine if a pair of random variables exhibits AI and the relative strength of dependence in the class. An alternative way to quantify the strength of association between the components with AI is through the coefficient of tail dependence $\eta \in (0,1]$ (Ledford and Tawn, 1996). The coefficient arises by approximating the joint survivor function of (X_i, X_j). Specifically, for $i, j \in A$, $i \neq j$, and as $u \to 1$

$$\mathbb{P}[F_{X_i}(X_i) > u, F_{X_j}(X_j) > u] = \mathcal{L}(1-u)(1-u)^{1/\eta}, \tag{1.3.2}$$

where $\mathcal{L}$ is a slowly varying function such that $\mathcal{L}(tx)/\mathcal{L}(x) \to 1$ as $x \to \infty$ for all positive constant t. An alternative way to obtain the coefficient of tail dependence η is

$$\eta = \lim_{u \to 1} \eta(u) = \lim_{u \to 1} \frac{\log(1-u)}{\log(\mathbb{P}[F_{X_i}(X_i) > u, F_{X_j}(X_j) > u])}$$

For pairs of random variables that exhibit AD, $\eta = 1$ and $\mathcal{L}(x) \nrightarrow 0$ as $x \to \infty$. Otherwise, the random variables exhibit AI, and η provides information on the strength of the association; if $\eta \in (0, 1/2)$, then the variables are negatively associated in the extremes; if $\eta \in (1/2, 1)$, then the variables are positively associated in the extremes. Exact independence is achieved if $\eta = 1/2$ and $\mathcal{L}(1-u) = 1$, otherwise we have near independence if $\eta = 1/2$ and $\mathcal{L}(1-u) \neq 1$. One can notice that η provides similar information to $\bar{\chi}$ since $\bar{\chi} = 2\eta - 1$.

Equation (1.3.2) can be extended to the multivariate setting; Eastoe and Tawn (2012) define the joint survivor function as

$$\mathbb{P}[F_i(X_i) > u : i \in A] = \mathcal{L}(1-u)(1-u)^{-1/\eta(A)-1}, \tag{1.3.3}$$

as $u \to 1$ and for $u \in (0, 1)$ and $\eta(A) \in (0, 1]$. The components of A exhibit AD if $\eta(A) = 1$ and $\lim_{u\to 1} \mathcal{L}(1-u) > 0$, otherwise they cannot be extreme simultaneously. Note, however, that subsets of A may still exhibit AD.

1.3.2 Conditional multivariate extreme value model

The conditional multivariate extreme value model (CMEVM) proposed by Heffernan and Tawn (2004) was the first model to provide a credible approach to data exhibiting both AI and AD. The CMEVM is not based on a multivariate distribution or process; rather, conditional on one variable being large, normalising functions are defined to control the rate of growth of all the other variables such that, after normalisation, the joint distribution of these "residuals" is non-degenerate.

To set up the model for the d-dimensional vector X, first define Y as the vector X after transformation onto standard Laplace margins. Note that the original model was defined on standard Gumbel margins (Heffernan and Resnick, 2007; Heffernan and Tawn, 2004), however, Keef et al. (2013) propose using standard Laplace margins to simplify the parameterisation of the dependence structure. Further, let $V_{|i} \coloneqq V \backslash \{i\}$, $X_{|i} \coloneqq \{X_j : j \in V_{|i}\}$, and $Y_{|i} \coloneqq \{Y_j : j \in V_{|i}\}$ denote the set V and vectors X, and Y excluding their ith element. The central modelling assumption is that there exist normalising functions $\{a_{j|i} : \mathbb{R} \to \mathbb{R}, j \in V_{|i}\}$ and $\{b_{j|i} : \mathbb{R} \to \mathbb{R}_+, j \in V_{|i}\}$, such that for any $i \in V$,

$$\left(\left\{ \frac{Y_j - a_{j|i}(Y_i)}{b_{j|i}(Y_i)} \right\}_{j \in V_{|i}}, \; Y_i - u_{Y_i} \right) \Bigg| \; Y_i > u_{Y_i} \xrightarrow{d} \left(\{Z_{j|i} : j \in V_{|i}\}, \; E \right), \tag{1.3.4}$$

with convergence as $u_{Y_i} \to \infty$. In the limit, the residual vector $Z_{|i} = \{Z_{j|i} : j \in V_{|i}\}$

is independent of the conditioning component Y_i and has a non-degenerate distribution, while the limit variable E follows a standard exponential distribution Heffernan and Resnick (2007). Consequently, inference can be undertaken separately on: (i) $Y_i \mid Y_i > u_{Y_i}$; (ii) the normalising functions; and (iii) the residuals $Z_{|i}$.

Step (i) is trivial, since the tail of $Y_i \mid Y_i > u_{Y_i}$ is standard exponential by limit (1.3.4). For (ii) the normalising functions, Heffernan and Tawn (2004) propose the form

$$a_{j|i}(y_i) = \alpha_{j|i} y_i, \quad b_{j|i}(y_i) = y_i^{\beta_{j|i}}$$

where, for Laplace margins, $\alpha_{j|i} \in [-1, 1]$ and $\beta_{j|i} \in (-\infty, 1]$. These flexible functions capture AD ($\alpha_{j|i} = 1$ and $\beta_{j|i} = 0$), complete independence ($\alpha_{j|i} = 0$), and AI (all other parameter combinations). While a class of models can be proposed for the normalising functions, there is no general class of distributions to model (iii) the residuals $Z_{|i}$. Heffernan and Tawn (2004) and Keef et al. (2013) use the working assumption that the components of $Z_{|i}$ are mutually independent and follow a Gaussian distribution. Consequently, estimation reduces to a regression problem that is computationally efficient and scalable to high dimensions, unlike most multivariate extreme value models. Conversely, the residuals are likely to be neither mutually independent nor Gaussian, meaning semi-parametric prediction is required, which limits the scalability to moderate dimensions.

1.4 Spatial extremes

Spatially referenced data is an example of multivariate data. A spatial process is denoted $\{X(s) : s \in \mathcal{S}\}$, where $\mathcal{S} \subset \mathbb{R}^2$. In this setting, we can significantly reduce the parameter space by exploiting the underlying spatial surface on which the data are collected. As a byproduct, we can also obtain stochastic simulations of the process at unobserved locations, thereby eliminating the need for post-inference interpolation typically required in multivariate modelling.

To achieve this, the summary measures χ_A and $\eta(A)$ in equations (1.3.1) and (1.3.3), respectively, need to be extended to the spatial setting (Huser and Wadsworth, 2022) to inform us how dependence in the tail changes with respect to a distance metric. Consider two distinct locations $s_1, s_2 \in \mathcal{S}$ such that $X(s_1) \sim F_1$ and $X(s_2) \sim F_2$. The pairwise spatial extension of χ_A is defined as

$$\chi(s_1, s_2) = \lim_{u \to 1} \chi_u(s_1, s_2) = \lim_{u \to 1} \mathbb{P}[X(s_1) > F_1^{-1}(u) \mid X(s_2) > F_2^{-1}(u)]. \tag{1.4.1}$$

The stochastic process is said to be asymptotically dependent if for any two sites the limit in equation (1.4.1) is positive, i.e. $\chi(s_1, s_2) > 0$ for all distinct $s_1, s_2 \in \mathcal{S}$. Otherwise, the process is deemed asymptotically independent. Similarly, the spatial extension of $\eta(A)$ can be defined by assuming as $u \to 1$ that

$$\mathbb{P}[X(s_1) > F_1^{-1}(u) \mid X(s_2) > F_2^{-1}(u)] \sim \mathcal{L}\{(1-u)^{-1}\}(1-u)^{1/\eta(s_1,s_2)-1}.$$

Note, this has the same interpretation as the bivariate form for η in Section 1.3.1. Processes observed on Euclidean spaces generally obey the rule that things close together are more similar than those that are far away (Wadsworth and Tawn, 2022). To explore how the extremal dependence decays with separation distance, we can obtain estimates of $\chi(s_1, s_2)$ for all $s_1, s_2 \in \mathcal{S}$. Plotting the estimates against the Euclidean separation distance

$$d(s_1, s_2) = ||s_1, s_2||$$

between the pairs, we should notice a smooth decay of $\chi(s_1, s_2)$ as $d(s_1, s_2)$ increase.

If the extremal dependence decays in a smooth manner with respect to Euclidean distance, then the dependence structure could be modelled using a Gaussian process with some correlation function $\rho(s_1, s_2) = C(d(s_1, s_2); \boldsymbol{\theta})$, such that $C(\cdot)$ is a function of distance $d(s_1, s_2)$ with respect to parameters $\boldsymbol{\theta}$. Thus, rather than having $d(d-1)/2$ dependence parameters in the correlation matrix, as is the case in the saturated multivariate model, we can approximate the correlation matrix with just a handful of parameters $\boldsymbol{\theta}$. Such a reduction in the parameter space provides large computational gains and allows the model to be fitted in higher dimensions.

Davison et al. (2012) and Huser and Wadsworth (2022) provide a comprehensive review of spatial extreme value models. The notion of an extreme event in a spatial context is unclear, and the precise definition leads to different modelling approaches. Given the block maxima approach in the univariate case, and the pointwise block maxima in the multivariate setting, a plausible approach is to use a max-stable process to model spatially indexed block maxima (Blanchet and Davison, 2011). However, such models have several limitations, including: inference is difficult due to the complicated likelihood, simulations are computationally expensive, the data is artificial, and the models can only capture AD. Rather than using artificial block maxima, the generalised Pareto distribution can be extended to consider generalised Pareto processes (Ferreira and de Haan, 2014) which condition on $\sup_{s\in\mathcal{S}}\{X(s)\}$ being large. Another natural extension is the r-Pareto process (de Fondeville and Davison, 2018) for the r-largest exceedances. Such models only result in AD over the entire spatial

domain, which is unrealistic if the spatial domain being considered is large. Thus, alternative constructions such as random scale mixtures (Huser et al., 2017) and spatial extensions of the CMEVM (Wadsworth and Tawn, 2022) have been developed.

All of the models mentioned are only applicable for modelling stochastic processes on Euclidean spaces. However, not all processes are observed on Euclidean spaces. For example, river discharges are measured on a river network. Such processes generally do not obey the consensus that things close together, in terms of their Euclidean distance, are more similar than those that are far away. Thus, the pairwise estimates of equations (1.4.1) may not decay smoothly with respect to Euclidean separation distance $d(s_1, s_2) = |s_1, s_2|$. Further, one cannot simply replace the Euclidean distance metric with a non-Euclidean distance metric in the correlation function $C(\cdot; \boldsymbol{\theta})$, as it may not yield a valid (positive definite) spatial dependence structure (Ver Hoef et al., 2006). Consequently, extending spatial extremes to non-Euclidean spaces is non-trivial and has consequently received little attention. While Asadi et al. (2015) have modelled river discharges in the upper Danube River basin, the model is a max-stable Brown-Resnick process and so is subject to the limitations discussed above. Extending spatial extremes to non-Euclidean spaces for models that can capture both AD and AI remains an open research question, which we contribute towards in this thesis.

1.5 Graph theory

Throughout this thesis, the data used in the applications is measured on a network. We briefly introduce how networks can be described and some of their main properties. A network can be represented by a graph $\mathcal{G} = (V, E)$, such that $V = \{1, \ldots, d\}$ is the set of distinct vertices/nodes and $E \subseteq \{\{j, k\} : j, k \in V, j \neq k\}$ is the set of edges/links. Consequently, within this thesis, we will only consider simple (no multiple edges between vertices), undirected graphs. In addition, we will only consider connected graphs (one can always find a path between each pair of vertices).

In Chapter 2, we are interested in the stochastic properties of the network. In particular, we are interested in modelling the degree distribution of the network using extreme value methods. The in-degree and out-degree distribution correspond to the number of edges entering and leaving a node, respectively. Since we are only considering undirected graphs, the two distributions will be equivalent.

In Chapters 3 and 4, the graph is no longer treated as the data, but rather, the graph is treated as the surface on which the data are observed. For Chapter 3, the locations where the data are observed are treated as the vertices, and the edges show the connections between

them. In the saturated case, where there exists an edge between every pair of vertices, the graph will be dense if the number of constituents in the graph is large. Consequently, it will be computationally expensive to fit and simulate from the resulting model, since there will be $d(d-1)/2$ parameters to be estimated in the correlation matrix. To address this, we exploit conditional independence statements. Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$, which follows a multivariate Gaussian distribution. For Gaussian graphical models, if $\{j, k\} \notin E$, then $X_j \perp\!\!\!\perp X_k \mid \boldsymbol{X}_{\setminus\{j,k\}}$. Whether the conditional independence is implied by the network on which the data are collected, or the "optimal" graphical structure that is learnt, these conditional independencies can substantially reduce the number of dependence parameters in the model, allowing for high-dimensional inference.

In Chapter 4, the network on which the data is measured is treated as the graph itself. Consequently, we require some measure of geodesic distance to determine the dependence between two locations on the network. For metric graphs (Bolin et al., 2024), the measure of distance is defined as the shortest path between any two locations, which is always valid since we only consider connected graphs. For our application to river discharges in the upper Danube River basin (Asadi et al., 2015), the measure of distance is simply the river distance between two locations. Consequently, the number of dependence parameters is drastically reduced to only a handful of parameters (see Section 1.4), which allows for potentially even higher-dimensional inference.

Bibliography

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, 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.

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, 5(3):1699–1725.

Bolin, D., Simas, A. B., and Wallin, J. (2024). Gaussian Whittle–Matérn fields on metric graphs. *Bernoulli*, 30(2):1611 – 1639.

Casey, A. and Papastathopoulos, I. (2023). Decomposable tail graphical models. *arXiv preprint arXiv:2302.05182*.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161 – 186.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

de Fondeville, R. and Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika*, 105(3):575–592.

Eastoe, E. F. and Tawn, J. A. (2012). Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika*, 99(1):43–55.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932.

Ferreira, A. and de Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717 – 1737.

Heffernan, J. E. and Resnick, S. I. (2007). Limit laws for random vectors with an extreme component. *The Annals of Applied Probability*, 17(2):537 – 571.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme

values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Huser, R., Opitz, T., and Thibaud, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatial Statistics*, 21:166–186.

Huser, R. and Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(1):e1537.

Joe, H. (1997). *Multivariate models and dependence concepts*. Chapman and Hall/CRC.

Keef, C., Papastathopoulos, I., and Tawn, J. A. (2013). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, 115:396–404.

Leadbetter, M. R. (1983). Extremes and local dependence in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65:291–306.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Murphy, C., Tawn, J. A., and Varty, Z. (2025). Automated threshold selection and associated inference uncertainty for univariate extremes. *Technometrics*, 67(2):215–224.

Northrop, P. J., Attalides, N., and Jonathan, P. (2016). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(1):93–120.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119 – 131.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10(1):33–60.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

Smith, R. L. (1987). Estimating tails of probability distributions. *The Annals of Statistics*, 15(3):1174 – 1207.

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–377.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv preprint arXiv:2102.00884*.

Ver Hoef, J. M., Peterson, E., and Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological statistics*, 13(4):449–464.

Wadsworth, J. and Tawn, J. (2022). Higher-dimensional spatial extremes via single-site conditioning. *Spatial Statistics*, 51:100677.

Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, 58(1):116–126.

Wadsworth, J. L. and Tawn, J. A. (2013). A new representation for multivariate tail probabilities. *Bernoulli*, 19:2689–2714.

Chapter 2

Modelling the Degree Distribution of Networks

Modelling the Degree Distribution of Networks

Abstract

Statistical modelling of the degree distributions of networks has received lots of attention since the introduction of the preferential attachment model (Barabási and Albert, 1999). However, most of these models assume a power-law distribution, which implies a narrow class of heavy-tailed behaviour. For those models that leverage ideas from extreme value theory, the majority have also restricted themselves to the heavy-tailed case. In addition, these statistical models tend to focus only on the large degrees, despite the low degrees exhibiting different behaviour. Such models are used to inform generative models that produce synthetic networks with certain properties, which are then limited due to the simplifying assumptions in the statistical modelling. Given these shortcomings, we follow the approach of Lee et al. (2024) and use a flexible mixture distribution, based on extreme value theory, for modelling the *entire* degree distribution. Our inference uses a likelihood framework, which requires the threshold(s) to be chosen in advance. To do this, we propose an extension to the automated threshold selection method of Murphy et al. (2025). We apply our proposed model to a range of networks to show its utility and flexibility.

2.1 Introduction

A network is a useful way of describing pairwise relationships in a complex system, with nodes (vertices) representing the constituents of the system and edges (links) characterising the interactions between them. Networks can be used to model systems in a broad range of disciplines, such as transport to portray infrastructure such as the London Underground (Domenico et al., 2014), biology to track the spread of viruses like COVID-19 (So et al., 2020), and sociology to show friendship interactions on social media platforms such as Facebook (Wang and Resnick, 2022).

Network science has two key themes. The first, which will be our focus, analyses features of real-world networks to understand their underlying structure and generating processes. The second then utilises this information to posit models that generate synthetic networks with the particular features analysed by the first (Barabási and Albert, 1999; Cirkovic et al., 2023; Erdös and Rényi, 1959; Watts and Strogatz, 1998). The two topics are thereby interlinked, with the former being essential to ensure that the latter does not impose unfounded or incorrect assumptions on the way the network is generated, which could lead to mistaken

conclusions being drawn (Khanin and Wit, 2006). One aspect of real-world networks that draws particular interest is the (in-)degree distribution, which is statistically modelled to understand the structure and growth of dynamics in networks, such as scientific citation (Steinbock et al., 2019). Constructing generative models that replicate the degree distribution is useful because it contains other information, such as the diameter (the average number of edges in the shortest path connecting any two nodes) and clustering coefficient (the proportion of realised connections among a node's neighbours compared with the number of all possible connections). Such properties may also be useful in explaining the underlying structure of the network (Wolf et al., 2002). For instance, networks with a small diameter generally have "hubs" that are well-connected, which is a feature of degree distributions that are heavy-tailed in nature.

Generally speaking, statistical models for the degree distribution assume the data follow a heavy-tailed distribution (Clauset et al., 2009). Ultimately, this limits the possible generative models that can produce synthetic networks with the same behaviour. More recently, extreme value methods have been used for modelling the degree distribution (Voitalov et al., 2019). However, this has also largely been restricted to the heavy-tailed domain.

The remainder of the section will therefore review the most common generating mechanisms and existing methods for statistical modelling of the degree distribution.

2.1.1 Generative models

One of the first generative models was the Erdős-Rényi random graph model (Erdös and Rényi, 1959). The model generates a network by first fixing the number of nodes n and then choosing $0 \leq m \leq \binom{n}{2}$, which are sequentially created to avoid repeated edges in the graph. More recently, the Watts and Strogatz small-world model (Watts and Strogatz, 1998), generates a network by connecting each of the n nodes with its l (or $l-1$ if l is odd) nearest neighbours. To induce randomness, each edge is rewired. Assume that $\{v_i, v_j\}$ is an edge in the graph, then this is rewired with probability p. If the edge is rewired, $\{v_i, v_j\}$ is replaced with $\{v_i, v_s\}$, where s is chosen with equal probability from the set $\{1, \dots n\} \backslash \{j\}$. Note that the rewiring may result in multiple edges and loops in the resulting graph. Let K be a discrete random variable, and k be realisations from said random variable, representing the degree distribution. Both models produce degree distributions that decay exponentially, i.e. $\mathbb{P}[K = k] \propto \exp(-k)$ for large k. Consequently, relatively well-connected nodes, termed "hubs", are almost non-existent in these synthetic networks.

de Solla Price (1965) notes that hubs are present in many real-world networks and suggests that power-law decay may be more appropriate to capture this feature. Power-law decay suggests the change in the frequency of a node's degree is relative to the change in the degree itself. Specifically, $\mathbb{P}[K = k] \propto k^{-\alpha}$, for large k and exponent parameter $\alpha > 1$. Networks with this behaviour are popularly termed "scale-free".

Although de Solla Price (1976) proposed the cumulative advantage process to generate synthetic networks with heavy tails, analysis of the degree distribution was not popularised until Barabási and Albert (1999) proposed the preferential attachment (PA) model, which builds on the former and produces networks where the degree distribution decays according to a power-law (Bollobás et al., 2001). The PA model generates a network by allowing the creation of new nodes and then the creation of new edges between the new and existing nodes. The probability of an edge being created with node i is equal to $k_i / \sum_j k_j$, where k_j is the current degree of node j. Consequently, already well-connected nodes have a higher probability of becoming even more connected. Following the proposal of the PA model, many real-world networks, including the World-Wide Web (Broder et al., 2000), metabolic networks (Jeong et al., 2000), computer virus epidemics (Pastor-Satorras, 2001), the number of sexual partners over a short period (Liljeros et al., 2001), and brain activity (Eguíluz, 2005), were subsequently deemed to arise according to the PA model. Their degree distributions are therefore deemed scale-free since $\mathbb{P}[K = k] \propto k^{-\alpha}$ for large k.

To more accurately capture the interactions of constituents in real-world networks, the PA model has received several adaptations. For instance, Bollobás et al. (2003) altered the model to produce directed networks, with different types of edges being created at each step. Further, Cirkovic et al. (2023) introduced the idea of reciprocity so that reciprocal edges can be formed in these directed networks.

2.1.2 Modelling the degree distribution

The power-law debate

While many argued for the ubiquity of the power-law in degree distributions and in favour of the PA model, others disagreed. For instance, Liljeros et al. (2001) highlighted that the left-tail of the degree distribution for the total number of sexual partners is non-linear, suggesting a power-law is only appropriate above some high threshold. This leads to ambiguity in the term scale-free; some authors mean the *entire* degree distribution follows a straight line on the log-log scale, while most only require linearity above some threshold, and some even ignore the randomness and/or the deviation of the largest degrees from the line. This imprecision in the term scale-free can mask the range of behaviour being exhibited in real-world networks,

which will hinder further developments of generative models.

Irrespective of what scale-free may mean, for a continuous random variable where power-law behaviour is only exhibited in the right tail, the excesses above a high threshold can be modelled using a Pareto distribution (or equivalently, the continuous power-law distribution). For some threshold $u > 0$, X follows a Pareto distribution if its probability density function (PDF) has the form

$$f(x \mid \alpha) = \frac{(\alpha - 1)}{u} \left(\frac{x}{u}\right)^{-\alpha}, \quad x > u, \tag{2.1.1}$$

with exponent parameter $\alpha > 1$. One can show that, just like the power-law, the distribution and survivor function have gradients of $-\alpha$ and $1 - \alpha$ on the log-log scale, respectively.

However, as noted by Clauset et al. (2009), the degree distribution is not a continuous random variable, meaning it should not be modelled using the Pareto distribution. Instead, its discrete counterpart, the Zipf (discrete Pareto or zeta) distribution, should be used. A discrete random variable R follows the Zipf distribution if its probability mass function (PMF) can be written as

$$p(r \mid \alpha) = \frac{r^{-\alpha}}{\zeta(\alpha, u)} = \frac{r^{-\alpha}}{\sum_{i=0}^{\infty}(u + i)^{-\alpha}}, \quad r \in \{u, u + 1, \dots\},$$

where $u > 0$, $\alpha > 1$, and $\zeta(\alpha, u)$ is the Hurwitz zeta function. Many authors now use the Zipf distribution for modelling (Jung and Phoa, 2021; Valero et al., 2022) the degree distribution.

Despite current advancements, initial statistical analyses of degree distribution were not very rigorous. For instance, Khanin and Wit (2006) noted that most networks are deemed scale-free by eyeballing or fitting a line of best fit to the degree distribution and/or survivor function (on the log-log scale). To make the analysis statistically robust, Khanin and Wit (2006) modelled the *entire* degree distribution using the Zipf distribution. They found that the power-law is inadequate for ten biological networks that were previously deemed scale-free. The differing conclusions may have arisen because Khanin and Wit (2006) modelled the entire degree distribution, while previous authors may have only suggested the power-law is present above a high threshold.

While the Zipf distribution can model the degree distribution, it has limitations. For instance, the model only fits well for heavy-tailed data because the exponent parameter must be greater than 1. This is not appropriate for all degree distributions, such as the number of collaborators on scientific papers which follows a power-law with exponential cut-off

(Newman, 2001), i.e. for large k

$$\mathbb{P}[K = k] \propto k^{-\alpha}\exp(-\lambda k), \tag{2.1.2}$$

such that $\alpha > 1$, $\lambda \geq 0$. This raises two concerns. First, we need statistical models that can capture a range of tail behaviour. Second, although the PA model appears to capture the behaviour of many real-world networks, it is unable to capture the behaviour of all networks, meaning other generative models are still required and should still be researched.

The debate on the ubiquity of the power-law culminated in Broido and Clauset (2019) claiming that scale-free networks are rare in practice. The authors present an algorithm to choose the best threshold above which the power-law is appropriate by minimising the Kolmogorov-Smirnov test statistic. Hypothesis tests are then performed to compare the power-law to alternative distributions. In most cases, a log-normal distribution is deemed the best fitting. The work received multiple criticisms. For instance, van der Hoorn et al. (2020) argue the assumption that the degree sequences are "pure" power-law is unrealistic as it is a limiting behaviour. Since the degree distribution and survivor function are taken at snapshots in time during the evolution of the network, it is possible that the network may not have reached its limiting behaviour, meaning the plots of the degree distribution and survivor function will not be exactly linear on the log-log scale. Consequently, the hypothesis tests conducted by Broido and Clauset (2019) were always going to favour the alternative hypothesis (the distribution does not follow the power-law).

The staunchest rebuttal came from Voitalov et al. (2019), who claimed scale-free networks are *not* rare in practice, although the results are not directly comparable due to the use of different datasets. The authors take inspiration from extreme value theory (EVT) and assume that the complementary cumulative distribution (or equivalently, survivor) function is regularly varying, i.e. $1 - \mathbb{P}[K \leq k] = \mathcal{L}(k)k^{-\alpha}$, where $\alpha > 0$ and $\mathcal{L}$ is a slowly varying function such that $\mathcal{L}(tx)/\mathcal{L}(x) \to 1$ as $x \to \infty$ for all positive constant t. Note that this is equivalent to saying the distribution is heavy-tailed, or slower than exponential according to Voitalov et al. (2019). Both the power-law and Pareto distributions are examples of regularly varying distributions. Voitalov et al. (2019) note that hypothesis tests are inappropriate when using regularly varying distributions since "the infinite number of degrees of freedom contained in the space of slowly varying functions makes the space of regularly varying distributions non-parametric". We disagree with the logic here since the Pareto and Zipf distributions are examples of regularly varying distributions that are also parametric. Thus, hypothesis tests can be conducted for these distributions. In the absence of hypothesis tests, they conclude that the degree distribution can be well approximated by a regularly varying distribution

only if the extreme value index $\xi = 1/(\alpha - 1)$ from multiple estimators (Hill (Hill, 1975), moment and kernel) are all positive. Despite taking inspiration from EVT, Voitalov et al. (2019) limit the inference to the heavy-tailed case by assuming regular variation. In addition, they model the discrete random variable using a continuous distribution by adding uniform noise to the data. While this approach does not lead to bias (see Section S2.2), a more elegant solution would be to use a discrete distribution.

Beyond the power-law

Arguably, the debate about whether networks are scale-free is counterproductive, and it is more important to know if a degree distribution is heavy-tailed or not (Holme, 2019; Stumpf and Porter, 2012). However, previous attempts to model the degree distribution, except Lee et al. (2024), have pre-specified the rate of tail decay to be heavy (Pareto/power-law), exponential (exponential/exponential cut-off), or light (Weibull), which limits the inference, and provides little information when the degree distribution does not satisfy the pre-specified rate of tail decay. To avoid such pre-specification, we utilise a model, similar to Lee et al. (2024), based on EVT to capture asymptotic behaviour that can range from light- to heavy-tailed. In addition, we elect to model the *entire* degree distribution, since there are many cases in which it is helpful to understand both the low- and high-degree behaviours, for example in community identity (Mehrabi et al., 2019) and percolation processes (Mannion and MacCarron, 2023). Thus, we combine traditional degree distribution models with models based on EVT to allow for potentially sub-asymptotic power-law behaviour. While some of the models proposed in Section 2.1.2 can model the entire degree distribution, most only apply above a high threshold. Very few studies have attempted to model the *entire* degree distribution beyond using a Zipf distribution. We provide a brief review of the more recent methods.

Chattopadhyay et al. (2021) recently introduced the modified Lomax distribution for modelling the degree distribution. However, this distribution is continuous and is not appropriate for modelling discrete random variables (see Section 2.4.1 for more details). Consequently, the parameter estimates, and possibly the fit, are likely to be biased, meaning the results are unreliable. Adopting a discretised version, obtained by integrating the PDF over unit intervals (Nakagawa and Osaki, 1975), would be more appropriate.

Previously, Chattopadhyay et al. (2014) utilised piecewise truncated geometric distributions for non-overlapping segments of the set $\{1, 2, \ldots\}$. The results suggest the mixture model provides a better fit to the tail compared to using a single power-law distribution above a high threshold. However, the number of segments must be carefully chosen to avoid under-

or over-fitting.

Rather than using a mixture distribution that segments the data, Jung and Phoa (2021) propose a mixture of weighted Zipf distributions. They lose some flexibility by restricting the exponent parameter α to be the same for each component. To try and recapture some of the lost flexibility, they allow each component to have different supports. The number of components and their support are estimated using an Expectation-Maximisation (EM) algorithm. The results show improvement over both a mixture of weighted power-law distributions and a single power-law distribution for the entire range. They also claim their model outperforms the generalised Pareto (GP) distribution and its discretised counterpart. However, the comparison is unreasonable since the distributions are fitted to the entire dataset rather than just the tail, meaning the EVT-based models are likely to be biased, since the asymptotic justification for these models is unlikely to be satisfied.

Rather than weighting the Zipf distribution, Valero et al. (2022) propose using the Zipf-polylog distribution. The model is extremely flexible as it can vary between the Zipf distribution and the polylog distribution to capture power-law and non-power-law-like behaviour, respectively. The polylog component essentially results in the distribution exhibiting power-law behaviour with exponential cut-off in equation (2.1.2). Although this model is exceptionally flexible, it has two drawbacks. Firstly, the distribution is unable to capture light-tailed data. Secondly, the model performs poorly on large networks (Valero et al., 2022, Figure 9) due to the changing behaviour of the degree distribution.

To overcome these shortcomings, Lee et al. (2024) adopt a mixture distribution for non-overlapping segments where the body components are modelled using the truncated Zipf-polylog distribution, and the tail is modelled using a discrete GP distribution. Overall, the model is flexible and can capture a range of behaviour in the bulk while also being able to capture light-, exponential-, and heavy-tails, since the rate of decay is a parameter to be estimated. To perform inference, Lee et al. (2024) adopt a Bayesian framework to allow for uncertainty in the estimated threshold to be incorporated into the model, as well as simple embedding of model selection. Given the model performance and flexibility, we adopt the same model proposed by Lee et al. (2024) but use an analogous frequentist framework. This provides an alternative inference procedure without the need for prior beliefs on the parameter estimates.

The remainder of this chapter is structured as follows. Section 2.2 introduces continuous and discrete mixture distributions, as well as the distributions we use for each component: the truncated Zipf-polylog and the discrete GP distributions. We also propose an extension to the automatic threshold method of Murphy et al. (2025) that selects the threshold by min-

imising the expected deviations in the quantile-quantile (QQ) plot to account for mixture distribution. Further, we propose a more principled search of the candidate thresholds that selects new thresholds to be tested based on their probability of reducing the distance metric, similar to the Bayesian optimisation approach adopted by Varty et al. (2021). Section 2.3 details how the inference is undertaken. Section 2.4 performs several simulation studies. First, we show how modelling discrete datasets with continuous distributions results in biased parameter estimates that may be masked by the model fit to highlight why discrete distributions are necessary for modelling discrete random variables. Second, we assess the performance of the proposed threshold selection method when the true underlying distribution and thresholds are known. Applications of the methods to real-world networks are presented in Section 2.5 before the chapter concludes with a discussion in Section 2.6.

2.2 Methodology

We aim to model the *entire* degree distribution using a mixture distribution. We therefore outline flexible distributions, the truncated Zipf-polylog and integer generalised Pareto distributions, which will be used to model the bulk (body) and tail, respectively. We then detail how these distributions can be incorporated into a mixture distribution. Inference will be performed using a likelihood framework, meaning the threshold(s) where one component ends and another begins must be chosen before fitting the model. Our proposed method for selecting these thresholds is presented in Section 2.2.4.

Note that we will assume the degrees are independent and identically distributed (IID) samples from some underlying distribution. While this may not be appropriate, as increasing the out-degree of one node inherently increases the in-degree of another node, it is common practice in the literature to assume independence between the degrees (Jung and Phoa, 2021; Lee et al., 2024; Valero et al., 2022; Voitalov et al., 2019). For the data used in the application, we only have degrees and frequencies, so we are unable to test how valid this assumption is. If we had unique identification numbers for each node, we could build a contingency table to test the independence assumption between the degrees. Accounting for the dependence between the degrees would be inherently complex and beyond the scope of our contribution.

2.2.1 Truncated Zipf-polylog distribution

The bulk of degree distributions can exhibit a range of behaviour, motivating the need for a flexible distribution to model it. One such distribution is the truncated Zipf-polylog(w, u, θ,

α) (TZP) distribution with PMF

$$p_{TZP}(z \mid w, u, \theta, \alpha) = \frac{z^{-\alpha}\theta^z}{\sum_{k=w+1}^{u} k^{-\alpha}\theta^k}, \quad z \in \{w+1, w+2, \ldots, u\}, \tag{2.2.1}$$

where $u > w \geq 0$ are integers, $\theta \in (0, 1]$. Note, $\alpha \in \mathbb{R}$ when $\theta \in (0, 1)$ *or* $u < \infty$, but $\alpha > 1$ when $\theta = 1$ *and* $u = \infty$.

Since the TZP distribution is the disjoint union of the Zipf-Mandelbrot(w, u, α) (ZM) and truncated polylog(w, u, θ, α) (TP), when $\theta = 1$ and $\theta \in (0, 1)$, respectively, it can smoothly transition between power-law and non-power-law behaviour (Valero et al., 2022). The distribution has many special cases, as depicted in Lee et al. (2024, Figure 2).

2.2.2 Integer generalised Pareto distribution

We briefly review the GP distribution from Section 1.2.2 before introducing its discrete counterpart. Recall for a *continuous* random variable X, the excess above a high threshold $X - u \mid X > u$ follows a GP distribution in the limit as $u \to x_F$, where $x_F = \sup\{x; F_X(x) < 1\}$ is the upper end-point of the underlying *absolutely continuous* distribution F (Pickands III, 1975). The GP distribution is the only distribution for which this holds.

The CDF $G_{u(x)}$ for $X - u \mid X > u$ is given in equation (1.2.5), while its PDF is given by

$$g_u(x \mid \sigma_u, \xi) = \begin{cases} \frac{1}{\sigma_u}\left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi-1} & \text{if } \xi \neq 0, \\ \frac{1}{\sigma_u}\exp\left[-\left(\frac{x-u}{\sigma_u}\right)\right] & \text{if } \xi = 0, \end{cases} \tag{2.2.2}$$

where $A_+ \coloneqq \max\{0, A\}$, $\sigma_u > 0$ is the threshold-dependent scale parameter, and $\xi \in \mathbb{R}$ is the shape parameter, with the case $\xi = 0$ being taken in the limit as $\xi \to 0$. To see that this generalises the Pareto distribution, recall that $\sigma_u = \sigma + \xi(u - \mu)$ (see Section 1.2.2). Now, when $\xi > 0$ and $\sigma = \xi\mu$, $g_u(x)$ is equivalent to the density of the Pareto distribution in equation (2.1.1) with $\alpha = 1/\xi + 1$.

One could approximate the degree distribution with a GP distribution, although this would be inappropriate since the degree distribution is a discrete random variable. Adopting such a practice may result in biased parameter estimates (see Section 2.4.1), since the underlying distribution function is not *absolutely continuous*. An alternative, proposed by Voitalov et al. (2019), is to add uniform noise to the integer-valued data before modelling it using the continuous GP distribution. Although this approach does not induce bias (see Supplementary

Material), it is not the most elegant solution given the ease with which discrete distributions can now be fitted.

A more sophisticated solution is to utilise the discrete counterpart of the GP distribution, the integer generalised Pareto (IGP) distribution. Discretisation is generally achieved by integrating the continuous density over unit intervals (a method was first proposed by Nakagawa and Osaki (1975) to discretise the Weibull distribution). The precise unit interval will determine the exact discretisation. Assume that the excess of a *continuous* random variable X above a threshold u follows a GP distribution, i.e. $X - u \mid X > u \sim \text{GP}(\sigma_u, \xi)$. Prieto et al. (2014) consider the random variable $Y = \lfloor X \rfloor$ which follows the IGP distribution such that

$$\mathbb{P}[Y = y \mid Y \geq \nu = \lfloor u \rfloor] = \begin{cases} G_\nu(y+1) - G_\nu(y) & \text{for } y \in \{\nu, \nu+1, \dots\}, \\ 0 & \text{otherwise.} \end{cases}$$

Note, Y is supported at ν and $G_u(\nu)$, the CDF of the *continuous* GP distribution in equation (1.2.5) evaluated at ν, is undefined. We therefore have to consider the excess of the random variable X above a threshold ν rather than u, i.e. the threshold in the continuous and discrete cases must be *identical* and an *integer*. Therefore, $X - \nu \mid X > \nu \sim \text{GP}(\sigma_\nu, \xi)$ where $\sigma_\nu = \sigma_u + \xi(\nu - u)$. Thus, Y follows the IGP distribution *under flooring* with threshold ν, scale parameter $\sigma_\nu > 0$ and shape parameter $\xi \in \mathbb{R}$. The shape parameter in the IGP distribution has the same value and interpretation as in its continuous counterpart.

Alternatively, Hitz et al. (2024) consider the random variable $Z = \lceil X \rceil$ which follows the IGP distribution *under ceiling* such that

$$\mathbb{P}[Z = z \mid Z > \nu = \lfloor u \rfloor] = \begin{cases} G_\nu(z) - G_\nu(z-1) & \text{for } z \in \{\nu+1, \nu+2, \dots\}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2.3}$$

This version is more widely used in the literature (Lee et al., 2024; Rohrbeck et al., 2018). Unless otherwise stated, we will assume the IGP distribution *under ceiling*.

2.2.3 Mixture distributions

As discussed in Section 2.1.2, the Zipf distribution, and flexible extensions such as the Zipf-polylog distribution, are inadequate for modelling the *entire* degree distribution. Given this, it is plausible that the low- and high-degree nodes have different underlying generating mechanisms, which we aim to capture using a mixture distribution. In the field of EVT, considerable research has been conducted on extreme value mixture models, which enable inference on both extreme and non-extreme events (André et al., 2024; Castro-Camilo et al.,

2019; Naveau et al., 2016; Stein, 2021).

As a side note, the term mixture distribution has several characterisations. The most widely used definition is a weighted sum of m distributions that belong to the same family. For example, Jung and Phoa (2021) use a mixture of truncated Zipf distributions to model the entire degree distribution. One can use a mixture of extreme value distributions in this manner, but this is currently restricted to the case where all components have non-negative shape parameters (Otiniano et al., 2017). It is possible to weight multiple distributions from different families. For example, Frigessi et al. (2002) weight a Weibull and a GP distribution to model insurance losses from industrial fires. However, the method is not widely used in EVT due to concerns that the fit from the bulk impacts the fit of the tail (and vice-versa), and that the limiting GP result may not hold since the GP component is applied over the entire support rather than above a high threshold. Thus, in EVT it is common to use piecewise distributions, which need not be from the same family (Lee et al., 2024; Zhao et al., 2010), to model non-overlapping segments of the data. Hereafter, the term mixture distribution will refer to this method, unless otherwise stated.

For mixture distributions, we use the definition of MacDonald et al. (2011). Assume that X is a *continuous* random variable. Further assume that below some threshold u, the bulk of the distribution has a parametric density function $h(\cdot \mid u, \boldsymbol{\lambda})$ with parameters $\boldsymbol{\lambda}$, while above the threshold, the tail has a parametric density function $g(\cdot \mid u, \boldsymbol{\beta})$ with parameters $\boldsymbol{\beta}$. The PDF of a *continuous* mixture distribution can then be defined as

$$f(x \mid u,\ \boldsymbol{\lambda},\ \boldsymbol{\beta}) = \begin{cases} \frac{1-\phi_u}{H(u|\boldsymbol{\lambda})} h(x \mid u,\ \boldsymbol{\lambda}) & \text{for } x \leq u, \\ \phi_u g(x \mid u,\ \boldsymbol{\beta}) & \text{for } x > u, \end{cases} \tag{2.2.4}$$

where $\phi_u = \mathbb{P}[X > u]$, and $H(u \mid \boldsymbol{\lambda}) = \int_0^u h(s \mid \boldsymbol{\lambda}) ds$ is a constant to ensures the bulk is supported on the correct range and the PDF integrates to 1. Assuming the tail of the distribution is the GP distribution, we replace $g(x \mid u, \boldsymbol{\beta})$ with equation (2.2.2).

One benefit of mixture models is that the threshold is more naturally embedded compared to the GP distribution. Consider the mixture model in equation (2.2.4). One could expect a smooth transition between the bulk and the tail of the distribution. For example, Carreau and Bengio (2009) model insurance claims, which generally have heavy/fat tails, using a Gaussian distribution for the body and a GP distribution for the tail. They impose continuity between them by equating the first and second derivatives at the threshold u. The constraints reduce the number of free parameters from five to three while simultaneously making the threshold a function of the parameters. Thus, no threshold selection is required, and we only have

one additional parameter compared to the GP distribution. This makes mixture models attractive when the entire distribution needs to be modelled, and the behaviour of the bulk and the tail differs.

To extend mixture models to integer-valued data, assume that $Z = \lceil X \rceil$ is a discrete random variable. For the tail, a natural choice is the IGP distribution in equation (2.2.3). For the bulk, any discrete distribution may be used. For modelling the degree distribution, the TZP distribution in equation (2.2.1) is a natural choice for the bulk due to its flexibility. Thus, a random variable Z that follows the TZP-IGP(w, u, θ, α, σ_u, ξ) distribution has PMF

$$p_2(z \mid \boldsymbol{\Theta}_2) = \begin{cases} (1-\phi_u)p_{TZP}(z \mid w, u, \theta, \alpha) & z \in \{w+1, w+2, \dots, u\}, \\ \phi_u \left[G_u(z \mid \sigma_u, \xi) - G_u(z-1 \mid \sigma_u, \xi)\right] & z \in \{u+1, u+2, \dots\}, \end{cases} \tag{2.2.5}$$

where $\boldsymbol{\Theta}_2 = (\theta, \alpha, \sigma_u, \xi)$, and $\phi_u = \mathbb{P}[Z > u]$. The case when $\theta = 1$ results in the ZM-IGP(w, u, 1, α, σ_u, ξ) distribution. Note that there is no additional normalising constant in the bulk component of equation (2.2.5) like in equation (2.2.4) since the TZP distribution is already truncated to exist on the correct domain.

The subscript "2" in $p_2(\cdot)$ denotes that there are two components in the mixture distribution. The distribution is not limited to two components and a discrete random variable Z is said to follow the three-component TZP-TZP-IGP(w, v, u, θ_1, α_1, θ_2, α_2 σ_u, ξ) mixture distribution if its PMF is

$$p_3(z \mid \boldsymbol{\Theta}_3) = \begin{cases} (1-\phi_{vu}-\phi_u)p_{TZP}(z \mid w, v, \theta_1, \alpha_1) & z \in \{w+1, w+2, \dots, v\}, \\ \phi_{vu}p_{TZP}(z \mid v, u, \theta_2, \alpha_2) & z \in \{v+1, v+2, \dots, u\}, \\ \phi_u \left[G_u(z \mid \sigma_u, \xi) - G_u(z-1 \mid \sigma_u, \xi)\right] & z \in \{u+1, u+2, \dots\}, \end{cases} \tag{2.2.6}$$

where $\boldsymbol{\Theta}_3 = (\theta_1, \alpha_1, \theta_2, \alpha_2, \sigma_u, \xi)$, and $\phi_{vu} = \mathbb{P}[v < Z \leq u]$. Similarly, when $\theta_1 = 1$ we obtain the ZM-TZP-IGP distribution, when $\theta_2 = 1$ we obtain the TZP-ZM-IGP distribution, and when $\theta_1 = \theta_2 = 1$ we obtain the ZM-ZM-IGP distribution.

One issue with the mixture distribution in equation (2.2.4), and its discrete counterparts in equations (2.2.5) and (2.2.6), is that a discontinuity can occur at the threshold if ϕ_u is unconstrained. We could impose a constraint in the first derivative, i.e. in the PMF (see Hu and Scarrott (2018) for more details). To constrain the model in this manner for the two-component mixture model, we would need to assume that the bulk is supported at $u+1$. This is something we do not explore, due to the unrealistic assumption that the bulk is valid outside its domain. A similar argument can be made for the three-component mixture model.

Therefore, ϕ_u and ϕ_{vu} will be unconstrained and estimated empirically; $\hat{\phi}_u = \sum_{i=1}^{n} \mathbb{1}\{z_i > u\}$ and $\hat{\phi}_{vu} = \sum_{i=1}^{n} \mathbb{1}\{z_i > v, z_i \leq u\}$.

2.2.4 Threshold selection

Using mixture distributions requires selecting the threshold where the bulk ends and the tail begins. Scarrott and MacDonald (2012) provide a comprehensive review of threshold selection methods for both *continuous* tail only and *continuous* extreme value mixture models (CEVMMs). However, these methods are not immediately applicable to *discrete* extreme value mixture models (DEVMMs). Thus, we need to explore alternative methods.

Rohrbeck et al. (2018) showed that the IGP under ceiling has threshold stability, meaning the same methods for choosing the threshold in the continuous tail only case (see Section 1.2.2) can be used to select the threshold in the discrete tail only case. However, these methods are not suitable for mixture models because they ignore the bulk when choosing the threshold, which is inappropriate since the bulk and the tail are not independent, particularly in the location of the threshold.

One method that could be adapted for DEVMMs is the quasi-likelihood approach of de Melo Mendes and Lopes (2004). They chose the proportion of points (and thereby the thresholds) in the left and right tails as the set that yields the highest log-likelihood value. This is similar to maximising the profile log-likelihood of the threshold(s), which, in our experience, is relatively flat for DEVMMs. While the method is computationally efficient, the threshold uncertainty is not accounted for, as the thresholds are treated as fixed once they have been selected. To overcome this limitation, the threshold(s) can be treated as parameter(s) to be estimated in a Bayesian framework. This was first done by Behrens et al. (2004) for CEVMMs and later adopted by Lee et al. (2024) for DEVMMs. In both cases, the priors are treated as independent, which is unlikely to be the case in practice due to the shared information between the bulk and tail.

In our model, inference will be performed using a likelihood framework (see Section 2.3 for details), which requires the threshold(s) to be chosen in advance. While simple methods to select the threshold(s) were investigated, they generally underestimated the true thresholds in our simulations and performed poorly when applied to real datasets. For example, Broido and Clauset (2019) fit an estimator for multiple thresholds and choose the optimal threshold as the one that minimises the Kolmogorov-Smirnov statistic. The statistic aims to minimise the distance between the empirical and model CDF (Dimitrova et al., 2020), which, due to the amount of mass in the bulk, results in the bulk being a non-parametric estimate at the

expense of the tail, and too low a threshold being selected. Alternatively, maximising the profile log-likelihood of the threshold(s) also results in underestimation due to the generally flat nature of the surface.

To ensure the fits of both the bulk and tail are considered, we elect for a bootstrapping procedure that will optimally estimate quantiles of the distribution. This extends the automatic threshold selection method, first proposed by Varty et al. (2021) and later generalised by Murphy et al. (2025), in the *continuous* tail only setting. The idea is to minimise the expected deviation of the QQ-plot, according to some distance metric, via non-parametric bootstrapping and Monte Carlo methods. One could argue the choice of metric is subjective, however, Varty et al. (2021) and Murphy et al. (2025) showed that the best metric should minimise the mean absolute error between the model and empirical quantiles. Extending the method to choose the threshold for the *discrete* extreme value mixture model is straightforward; all that is required is to change the likelihood and quantile functions involved. Our contribution lies in making the procedure more computationally efficient.

Gradient-based optimisation procedures are inappropriate for minimising the metric, since the Monte Carlo procedure induces local roughness into the space (Varty et al., 2021). Therefore, Murphy et al. (2025) choose the set of candidate thresholds to be equally spaced between some quantiles of the data. For our purposes, thresholds must be integers, as candidates between integers will provide indistinguishable differences. Testing every unique integer between some quantiles is computationally expensive for larger networks and is computationally infeasible for the three-component mixture model when pairs of thresholds need to be tested. Varty et al. (2021) avoids testing every possible threshold by using a Bayesian optimisation procedure to search the set of candidate thresholds. We also search the space in a principled manner by utilising a multi-arm bandit approach (Slivkins, 2019) described below.

We start with a set of candidate thresholds. Next, choose k candidates and determine the distance metric. The untested candidates are assigned a probability that is proportional to the weighted sum of distance metrics of their two closest neighbours, in terms of Euclidean distance, that have been tested. The probability up/down weights candidates in the vicinity of thresholds with low/high distance metrics. Using the probabilities, k new candidates are chosen and tested. The process repeats until we observe s consecutive trials without reducing the distance metric.

The hyper-parameters k and s will depend on the size of the network and need to be sensibly tuned by the user; if k and/or s are too small, then the method will not search enough of the space and result in a sub-optimal threshold; conversely, if k and/or s are too large, then too many candidates will be tested resulting in a large computational cost. Provided that

Algorithm 2.1 Proposed threshold selection procedure

Initialise $\boldsymbol{z}$, C_u, $B \in \mathbb{N}$, $m \in \mathbb{N}$, $k \in \mathbb{N}$, $s \in \mathbb{N}$
Let $\boldsymbol{p}$ be a vector equally spaced probabilities such that $p_l = \{l/(m+1)\ ;\ l = 1, \ldots, m\}$
Let $\boldsymbol{d}$ be an empty vector of distances
Set $s^* = 0$, and $d^* = 10^{10}$
Select k equally spaced candidates from C_u and denote them $C_u^{(i)}$ for $i = 1, 2, \ldots, k$
Let C_u^* and $C_u^{'} = C_u \backslash C_u^*$ be the sets of tested and untested candidates, respectively
while $s^* \leq s$ **do**
  for i `in 1 to` k **do**
    for j `in 1 to` B **do**
      Obtain a non-parametric bootstrap sample $\boldsymbol{z}^{(j)}$ from $\boldsymbol{z}$
      Fit the model to $\boldsymbol{z}^{(j)}$ with thresholds $C_u^{(i)}$ and obtain $\hat{\boldsymbol{\Theta}}^{(j)}$
      Calculate $q^{(j)}(\boldsymbol{p}) = \frac{1}{l}\sum_{l=1}^{m} \mid \hat{Q}^{(j)}\big(p_l; \hat{\boldsymbol{\Theta}}^{(j)}\big) - Q^{(j)}\big(p_l\big) \mid$; the mean absolute error between the model and empirical quantiles of $\boldsymbol{z}^{(j)}$
    end for
    Append $\frac{1}{B}\sum_{j=1}^{B} q^{(j)}(\boldsymbol{p})$ to $\boldsymbol{d}$
  end for
  if $\min\{\boldsymbol{d}\} < d^*$ **then**
    Set $d^* = \min\{\boldsymbol{d}\}$, and $s^* = 0$
  else
    Set $s^* = s^* + 1$
  end if
  Append $C_u^{(i)}$, for $i = 1, \ldots, k$ to C_u^* and determine $C_u^{'}$
  Let n^* and n' be the number of tested and untested candidates, respectively
  for i `in 1 to` n' **do**
    for j `in 1 to` n^* **do**
      Let e_j be the Euclidean distance between $(C_u^*)^{(j)}$ and $(C_u^{'})^{(i)}$
    end for
    Let $\boldsymbol{r}$ be the indices of the two smallest value of $\boldsymbol{e}$
    Let $w_i = (e_{r_1} + e_{r_2}) \,/\, (d_{r_1} e_{r_1} + d_{r_2} e_{r_2})$
  end for
  Let $g_i = w_i / \sum_{j=1}^{n^{'}} w_j$
  Sample k thresholds from $C_u^{'}$ with probabilities $\boldsymbol{g}$
end while

k and s are sufficiently large, the algorithm ensures that the space is sufficiently explored without extensively searching areas unlikely to reduce the distance metric. Full details of the procedure, including the distance metric, are provided in Algorithm 2.1 where C_u is the set of candidate thresholds, B is the number of bootstraps, and m is the number of probabilities.

Note that while the proposed approach has links to gradient-based optimisation procedures, the exploration element of the multi-arm bandit approach provides a subtle difference. The

proposed approach has a non-zero probability of exploring a new region where a new minimum may be found. Thus, provided that k and s are sufficiently large, the algorithm will find the global minimum. However, this is not the case for gradient-based methods, which may converge to a local minimum.

In step 10 of Algorithm 2.1, we non-parametrically bootstrap the data. Parametric bootstrapping could be employed, but our simulation studies revealed that this generally results in a higher threshold being chosen compared to the true threshold. Consequently, estimation of the bulk is poor, and there is additional variability in the tail, both of which are undesirable. Furthermore, our non-parametric bootstrap treats the number of points in the tail as random. While the number of points could be fixed, the random approach is computationally simpler and allows for uncertainty in the proportion of points that exceed the threshold.

Once the threshold selection procedure has been performed, the threshold(s) are subsequently treated as fixed. To account for threshold uncertainty, the data would need to be bootstrapped, and the threshold selection procedure repeated for each bootstrapped sample. This is not explored here due to the computational costs associated with accounting for the threshold uncertainty.

For completeness, there are two computational bottlenecks with the Algorithm 2.1. Firstly, for every candidate threshold, we must take a large number of bootstrap samples, refit the model to each sample, and determine the distance metric (steps 8 - 15). Secondly, we need to recalculate the weights (steps 23 - 29) in each iteration. For the threshold selection simulation study in Section S2.3.2, where data are generated from the geometric distribution, these steps account for, on average, 99.977% and 0.022% of the total time taken per iteration, respectively.

To reduce the first computational bottleneck, one could argue that if the model is a poor fit for the original data, then it will provide a poor fit for the bootstrapped sample. Thus, we could avoid refitting and use parameter estimates from the model fitted to the original data. However, this will have a larger impact when there are fewer points in the bulk/tail due to larger parameter uncertainty. The second computational bottleneck seems small in comparison. However, this element becomes more prominent for larger (in terms of the number of unique values) datasets. To speed this component up, one could store the weights from each iteration and only update the weights for those where at least one closer candidate threshold has been tested compared to the previous iteration.

2.3 Inference

Inference will be performed using a likelihood framework, assuming that the thresholds are known. Due to the discrete nature of degree distributions, there will be multiple observations with the same integer value. For an IID sample of size n, there will be $m \leq n$ unique values $\boldsymbol{z} = (z_1, \ldots, z_m)$ with counts $\boldsymbol{c} = (c_1, \ldots, c_m)$ such that $\sum_{i=1}^{m} c_i = n$. Assuming that Z follows the two-component mixture distribution in equation (2.2.5), the log-likelihood is

$$\begin{aligned} l_2(\boldsymbol{\Theta}_2; \boldsymbol{z}, \boldsymbol{c}) = (n - n_u)\log(1 - \phi_u) + n_u\log(\phi_u) + \\ \sum_{i=1}^{m} c_i \log(p_{TZP}(z_i; w, u, \theta_1, \alpha_1))\mathbb{1}\{w < z_i \leq u\} + \\ \sum_{i=1}^{m} c_i \log(G_u(z_i; \sigma_u, \xi) - G_u(z_i - 1; \sigma_u, \xi))\mathbb{1}\{z_i > u\}, \end{aligned} \tag{2.3.1}$$

where $\boldsymbol{\Theta}_2 = (\theta_1, \alpha_1, \sigma_u, \xi)$, $n_u = \sum_{i=1}^{m} c_i \mathbb{1}\{z_i > u\}$ is the number excesses above u, and $\hat{\phi}_u = n_u/n$. Note, ϕ_u is not a free parameter as w and u are assumed to be known.

The log-likelihood for the three-component mixture model in equation (2.2.6) is

$$\begin{aligned} l_3(\boldsymbol{\Theta}_3; \boldsymbol{z}, \boldsymbol{c}) = (n - n_{vu} - n_u)\log(1 - \phi_{vu} - \phi_u) + n_{vu}\log(\phi_{vu}) + n_u\log(\phi_u) + \\ \sum_{i=1}^{m} c_i \log(p_{TZP}(z_i; w, v, \theta_1, \alpha_1))\mathbb{1}\{w < z_i \leq v\} + \\ \sum_{i=1}^{m} c_i \log(p_{TZP}(z_i; v, u, \theta_2, \alpha_2))\mathbb{1}\{v < z_i \leq u\} + \\ \sum_{i=1}^{m} c_i \log(G_u(z_i; \sigma_u, \xi) - G_u(z_i - 1; \sigma_u, \xi))\mathbb{1}\{z_i > u\}, \end{aligned} \tag{2.3.2}$$

where $\boldsymbol{\Theta}_3 = (\theta_1, \alpha_1, \theta_2, \alpha_2, \sigma_u, \xi)$, $n_{vu} = \sum_{i=1}^{m} c_i \mathbb{1}\{v < z_i \leq u\}$, and $\hat{\phi}_{vu} = n_{vu}/n$.

Inference for log-likelihoods (2.3.1) and (2.3.2) is performed using numerical optimisation. However, this can be reduced to maximising the profile log-likelihood for the components since the data are segregated into non-overlapping segments.

Note, since the thresholds are treated as known, the likelihood framework is quick and unbiased. Other fitting procedures, such as the Expectation-Maximisation (EM) algorithm, could be used for inference in these circumstances. However, if the thresholds were unknown, the EM algorithm would not be appropriate for inferring the thresholds and model parameters because there is an inherent overlap of information between the bulk and tail, particularly in the location of the thresholds.

2.4 Simulation study

In this section, we perform two simulation studies. The first details how modelling discrete data with a continuous distribution can yield biased parameter estimates, which are masked by the model fit. The second assesses the performance of the threshold selection procedure proposed in Algorithm 2.1.

2.4.1 Why are discrete distributions necessary?

Hitz et al. (2024) explain that modelling discrete data with a continuous distribution can result in biased parameter estimates due to ties in the data, which can impact the model fit and lead to erroneous conclusions. Thus, such practices are not advisable. If it cannot be avoided, a clear warning should be provided to the reader that the distribution is assumed to model data on a more refined scale than is available. However, in the network science literature, the degree distribution, a discrete random variable, is sometimes modelled using a continuous distribution without such warnings (Chattopadhyay et al., 2021; Jeong et al., 2000). Furthermore, they claim their model provides a better fit than those based on existing methods. Given the bias, it is unclear whether these claims are accurate.

When assessing the tail of a distribution, approximating a discrete random variable with a continuous distribution comes with additional complexities. This is because the maximum domain of attraction (DoA) is not necessarily preserved when we discretise a distribution in the exponential and light-tailed cases. For example, the exponential distribution belongs to the DoA of the Gumbel distribution, while its discrete counterpart, the geometric distribution, does not (Shimura, 2012). As such, using a continuous distribution could, in theory, result in the incorrect rate of tail decay, which has major consequences when we extrapolate beyond the range of the observed data. For network data, extrapolation can be used for predicting the size, in terms of the number of nodes, and the connectivity of a network in the future, which can be used for planning purposes. For example, for the "com-amazon" dataset analysed in Section 2.5, if the rate of tail decay is underestimated, then we will not extrapolate as many high-degree nodes, which could lead to outages if the servers cannot manage the traffic on the website. Conversely, if the rate of tail decay is overestimated, then we will extrapolate too many high-degree nodes, which could lead to wasted resources. In either case, the company loses money due to poor extrapolation. The remainder of this subsection outlines the dangers of using a model that does not correspond to the coarseness of the available data.

Assume that X follows a GP distribution with threshold $u = 10$, scale parameter, $\sigma_u = 20$

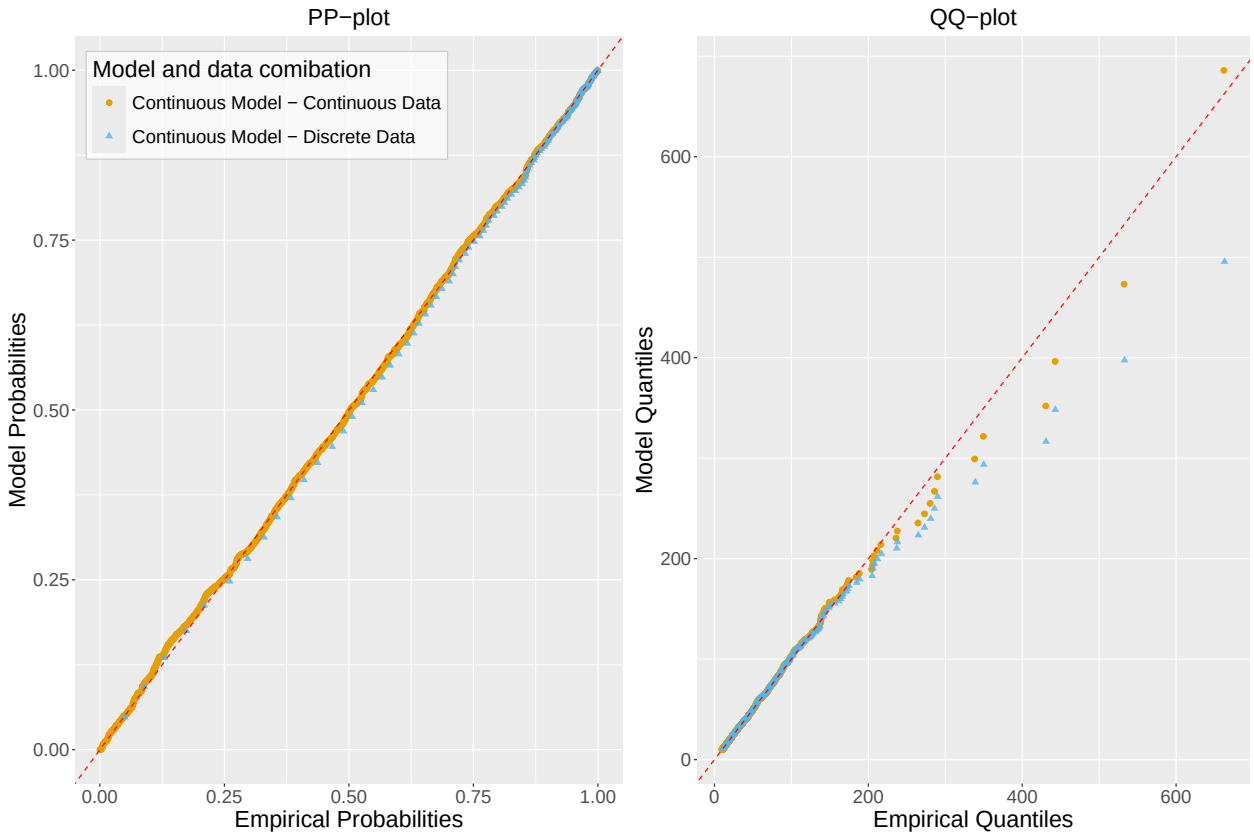

Figure 2.1: PP- (left) and QQ-plot (right) when the GP distribution is used to model the continuous (orange circles) and discrete (blue triangles) dataset. The red dashed line represents the $y = x$ line.

and shape parameter $\xi = 0.3$, i.e. $X \sim \text{GP}(10, 20, 0.3)$, and $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ are n IID realisations from X. We generate 1000 replicates of $n = 1500$ realisations from X such that $x_i^{(j)}$ is the ith realisation from the jth replicate, for $i = 1, \ldots, 1500$, and $j = 1, \ldots, 1000$. Assume that $Y = \lceil X \rceil$, i.e. Y follows an IGP *under ceiling* with the same parameters. To obtain samples from the IGP, it is sufficient to take the ceiling of our continuous realisations, i.e., $y_i^{(j)} = \lceil x_i^{(j)} \rceil$.

We model each continuous dataset, $\boldsymbol{x}^{(j)}$, using a GP with a threshold $u = 10$. For the discrete datasets, we fit both the GP and the IGP using a threshold of 10. Figure 2.1 shows the PP- and QQ-plots for a randomly selected replicate. For readability, we have excluded when the discrete dataset is modelled by the IGP distribution since the probabilities/quantiles are almost identical to those when the continuous dataset is modelled using the GP distribution. In addition, to better illustrate the goodness-of-fit (or lack thereof), we have only shown the points where there are jumps in the empirical CDF in the PP-plot for the discrete dataset. This practice will be employed throughout when modelling discrete data.

The model diagnostics in Figure 2.1 are somewhat deceptive. Although both models have good agreement with the $y = x$ line, there are issues with the fit. In the region 0.25 - 0.925 of the PP-plot, the continuous model for the discrete dataset consistently underestimates the empirical probabilities. Conversely, the continuous model for the continuous dataset exhibits both under- and over-estimation, which is what we would expect from a well-fitting model. Another point is that the probabilities from the continuous model for the discrete dataset in the region 0.925 - 1 are very close to the $y = x$ line due to 1 being the edge of the support. However, this masks the lack of fit indicated by the QQ-plot, where systemic underestimation is apparent in the largest quantiles when compared to the quantiles from the continuous model for the continuous dataset. This also highlights why you may wish to use a distance metric based on deviations in the QQ-plot rather than the PP-plot in Algorithm 2.1, as the former will penalise the lack of fit more. Therefore, even though the model fit does not appear bad when the discrete data is modelled using a continuous distribution, it is clear that the model is not appropriate when we closely inspect the fit.

The reason for the lack of fit is highlighted in Figure 2.2, which compares the parameter estimates from each model fit. The left panel compares the parameter estimates when the GP distribution models the continuous and discrete datasets. For the scale parameter (top), the estimates are parallel to, but above, the $y = x$ line, while for the shape parameter (bottom), the estimates are parallel to, but below, the $y = x$ line. Since the estimates are parallel to the $y = x$ line, it suggests the GP distribution for the discrete datasets is the biased model. Similar conclusions can be drawn when assessing the centre panels of Figure 2.2, which compares the estimates from the GP and IGP distributions for the discretised datasets. The right panels of Figure 2.2 compare the parameter estimates when the GP distribution models the continuous datasets and the IGP distribution models the discrete datasets. We observe almost perfect alignment of the parameter estimates with each other and the $y = x$ line, and they are evenly distributed around the true parameter values (the blue squares). This suggests that the IGP distribution is sufficient for handling the coarseness of the discretised data in a way that the GP distribution cannot.

In the left and centre panels of Figure 2.2, we observe that the scale parameter of the GP distribution exhibits positive bias while the shape parameter exhibits negative bias. This is expected since the two parameters are negatively correlated. A natural question would be, why is the bias not the other way around? Hitz et al. (2024) notes that ties in the data cause the bias, and we conjecture that the ties cause the continuous model to underestimate the shape parameter. Although not shown here, if the simulation study is repeated with $\sigma_u = 200$ and $\xi = 0.3$, the resulting discrete samples have few repeated values and are "approximately" continuous. Consequently, the parameters are still biased in the same direction, but the bias

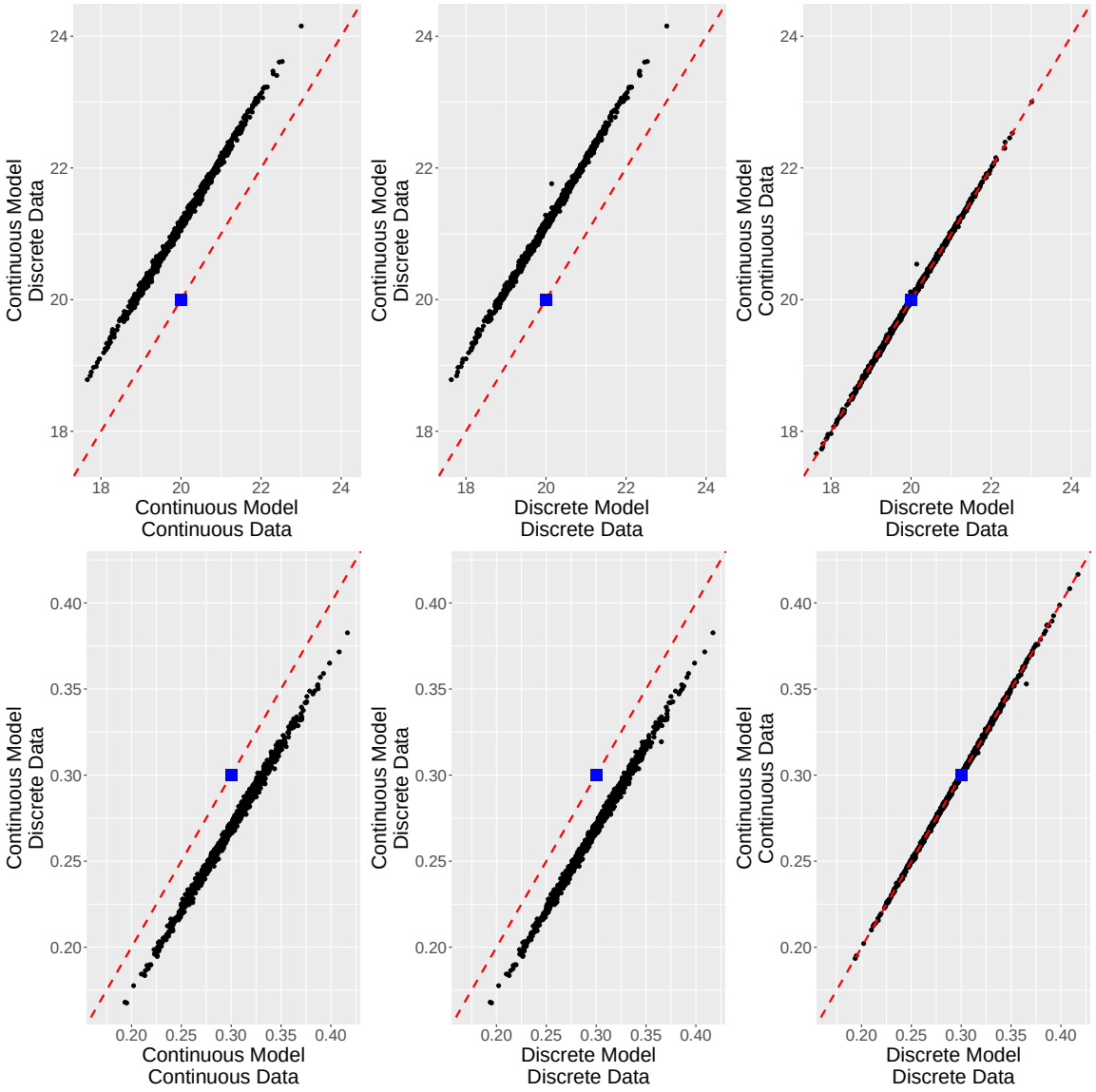

Figure 2.2: Scatter plots of 1000 parameter estimates (top - scale, bottom - shape). The left panels compare the maximum likelihood estimates (MLEs) when the GP distribution is used to model both continuous and discrete datasets. The centre panels compare the MLEs when the GP and IGP distributions are used to model the discrete datasets. The right panels compare the MLEs when the GP distribution models the continuous datasets and the IGP distribution models the discrete datasets. The red dashed lines represent the $y = x$ line. The blue square corresponds to the true value of the parameter.

is significantly reduced.

This example only shows bias from the GP distribution for this set of parameter values; however, similar results are obtained when alternative sets of parameters are used. In addition, the standard errors of the parameter estimates will also be affected by this bias, but this will not be investigated here. Further, the phenomenon of bias due to model mis-specification is

not restricted to the GP distribution. Similar commentary for the exponential and geometric distributions, where the fit is drastically impacted by model mis-specification, is provided in the Supplementary Material.

To summarise, although using the GP distribution to model the discrete datasets does not drastically impact the overall fit (see Figure 2.1), we obtain biased parameter estimates when the model is mis-specified (see Figure 2.2). Considering we do not know the true parameter values for real-world data, using a continuous model for discrete data could lead to erroneous conclusions if we extrapolate far beyond the range of the observed data. Therefore, we strongly advise against such practices and recommend that *continuous* distributions are only used to model *continuous* data, and *discrete* distributions are only used to model *discrete* data.

2.4.2 Threshold selection performance

In this subsection, we assess the performance of our threshold selection method on data generated from the TZP-ZM-IGP distribution. Specifically, we simulate 100 replicates of size 5000 from the TZP-ZM-IGP with $\boldsymbol{\Theta}_3 = (0.95, 0, 1, 1.7, 50, 0.3)$, and $w = 0$, $v = 30$ and $u = 130$. We then perform the threshold selection procedure in Algorithm 2.1 with $k = s = 10$ on each replicate to determine the optimal thresholds. Note, 100 replicates is low, and we accept this will lead to larger variability in our bootstrapped estimates, but 100 replicates is sufficient to assess the performance of the algorithm without incurring excessive computational costs. We perform this for both the TZP-TZP-IGP and the TZP-ZM-IGP to assess the impact of model mis-specification on the procedure. After obtaining the thresholds, we numerically optimise the log-likelihood (2.3.2), treating the thresholds as fixed.

Figure 2.3 displays boxplots of the selected thresholds and subsequent parameter estimates, over the 100 replicates. Interestingly, the TZP-TZP-IGP provides an unbiased threshold for v and a slightly positive biased threshold for u, while the converse applies for the TZP-ZM-IGP. Although the TZP-TZP-IGP is unbiased for v, there are many outliers, which raise questions about the suitability of the model. Since the TZP-ZM-IGP overestimates v, it is unsurprising that there is a small positive bias in θ_1 and α_1. However, the bias is comparable with that produced by the TZP-TZP-IGP, suggesting even a small bias in the threshold v has little impact on the subsequent parameter estimates. Moving on to the second component, the TZP-TZP-IGP exhibits negative bias, which is expected since the model is mis-specified in this component, and is possibly exacerbated by the component modelling more data than it should (since u is overestimated). The TZP-ZM-IGP is unbiased in the second component, which is possibly aided by modelling less data than it ought to (since v

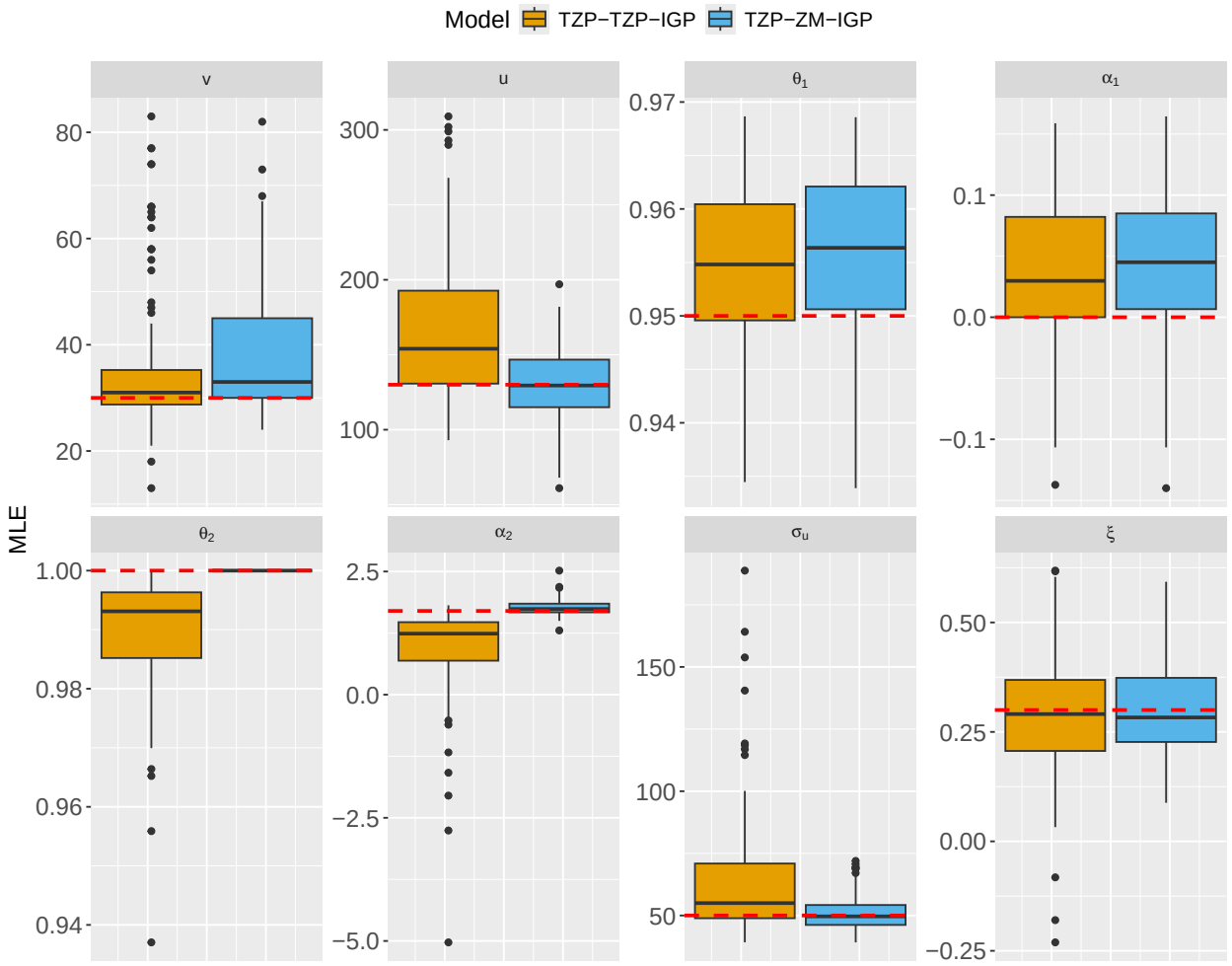

Figure 2.3: Boxplots of estimated thresholds and parameters of the TZP-TZP-IGP (orange) and TZP-ZM-IGP (blue) over 100 replicates. The thresholds and parameters are v, u, θ_1, α_1, θ_2, α_2, σ_u, and ξ from top left to bottom right. The red dashed horizontal line in each panel denotes the true threshold/parameter value.

is overestimated). For the IGP component, the scale and shape are unbiased for the TZP-ZM-IGP since u is unbiased. For the TZP-TZP-IGP, the shape parameter is unbiased, while the scale parameter is positively biased. This is expected since u is positively biased and the scale parameter is threshold-dependent, while the shape parameter is not. Overall, these results suggest the threshold selection procedure is performing well and that small biases in the selected threshold have a negligible impact on the parameter estimates, provided the model is correctly specified.

To visualise the utility of the threshold selection procedure, Figure 2.4 provides the output from the algorithm for a single replicate. The points depict the tested candidates, which are coloured by their subsequent distance metric. In both panels, few points are tested when

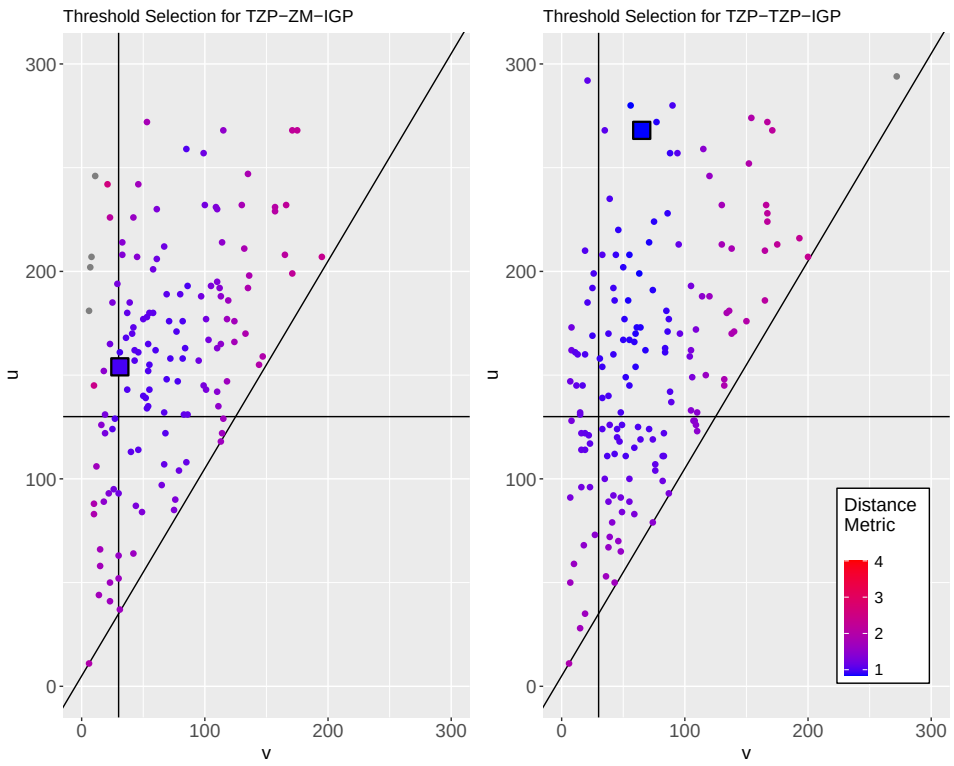

Figure 2.4: Scatter plots of tested candidate thresholds for the TZP-ZM-IGP (left) and TZP-TZP-IGP (right) using the threshold selection procedure in Algorithm 2.1. The vertical and horizontal black lines denote the true value of v and u, respectively. The diagonal black line represents the $u = v + 5$ line (the lower bound for v). The colour of the points details the value of the distance metric. The large square with a black outline is the threshold that minimises the distance metric, amongst those tested, in each case.

both v and u are large. This is because the distance metrics of the tested candidates in this region are large, suggesting these combinations are unlikely in practice. Similar findings are made when $v < 50$ and $u > 200$, and when both v and u are less than 50. Conversely, many candidates are tested close to the true thresholds, and where points have a lower distance metric (bluer in colour).

For the TZP-ZM-IGP (left panel), the procedure converges quickly. However, the optimal threshold, denoted by the large square with a black outline, slightly overestimates u despite testing a candidate very close to the truth. The overestimation might be due to simulation uncertainty, and increasing the number of bootstrapped samples B, currently set at 200, in Algorithm 2.1 may result in a threshold closer to the truth being selected. For the TZP-TZP-IGP, more candidates are tested, and the chosen threshold overestimates both v and u. Better thresholds may be obtained in both cases by increasing k and/or s; however, this is not explored further. Overall, the procedure is performing as expected by sparsely testing

regions that are unlikely to reduce the distance metric, and exploring areas that have already yielded low distance metric and thereby low deviation QQ-plots. The proposed method is also computationally more efficient than testing all candidates (every pair of integers), as, on average, only 0.7% of all possible candidates are tested. Additional examples assessing the performance of the threshold selection procedure in an edge case and for data not generated from a mixture distribution are provided in the Supplementary Material.

Despite the models overestimating at least one of the thresholds, both models fit the data well. Figure 2.5 shows the QQ-plot (left) and the survivor function (right), both on the log-log scale, for the single replicate. The QQ-plot is presented on the log-log scale to assess both the lower- and upper-tail fits. We observe that both models capture both tails well, with close agreement to the $y = x$ line. From the survivor function plot, it is clear that the TZP-ZM-IGP is the better-fitting model as the 95% confidence interval, obtained over 200 bootstraps with the thresholds treated as fixed to those previously estimated, is significantly narrower and better captures the upper-tail with the trajectory of the band being more aligned with the data. This is likely due to u being better estimated, which in turn results in σ_u being estimated closer to its true value.

Using model selection criteria, the AIC and BIC correctly choose the TZP-ZM-IGP 62 and 92 times out of 100, respectively. The AIC prefers more complicated models, which explains why it prefers the TZP-TZP-IGP more often. One can also argue that the TZP-TZP-IGP is degenerating to the TZP-ZM-IGP, since $\hat{\theta}_2 \approx 1$ and $\hat{\alpha}_2 \geq 1$ for the majority of the replicates (see Figure 2.3) for the TZP-TZP-IGP. Coupled with the visual diagnostics, it is clear that the TZP-ZM-IGP is the better-fitting model, however, this needs to be methodically deduced.

2.5 Application

We now assess the fit of our proposed model to various datasets. The majority of the datasets are undirected, unweighted networks from the KONECT database (Kunegis, 2013). While Voitalov et al. (2019) analyses 115 such networks, we take a smaller sample of 17 networks that omit any degree distributions that are too small for a mixture distribution to fit well to, and those with clear, possibly artificial, jumps in the degree distribution. As per Lee et al. (2024), we also analyse some alternative datasets available in the `poweRlaw` package (Gillespie, 2015) in `R` (R Core Team, 2025) that, while not being degree distributions, have been analysed to assess whether they follow the power-law or not. Specifically, they are two sets of casualty numbers in armed conflict ("us-americans" and "native-americans")

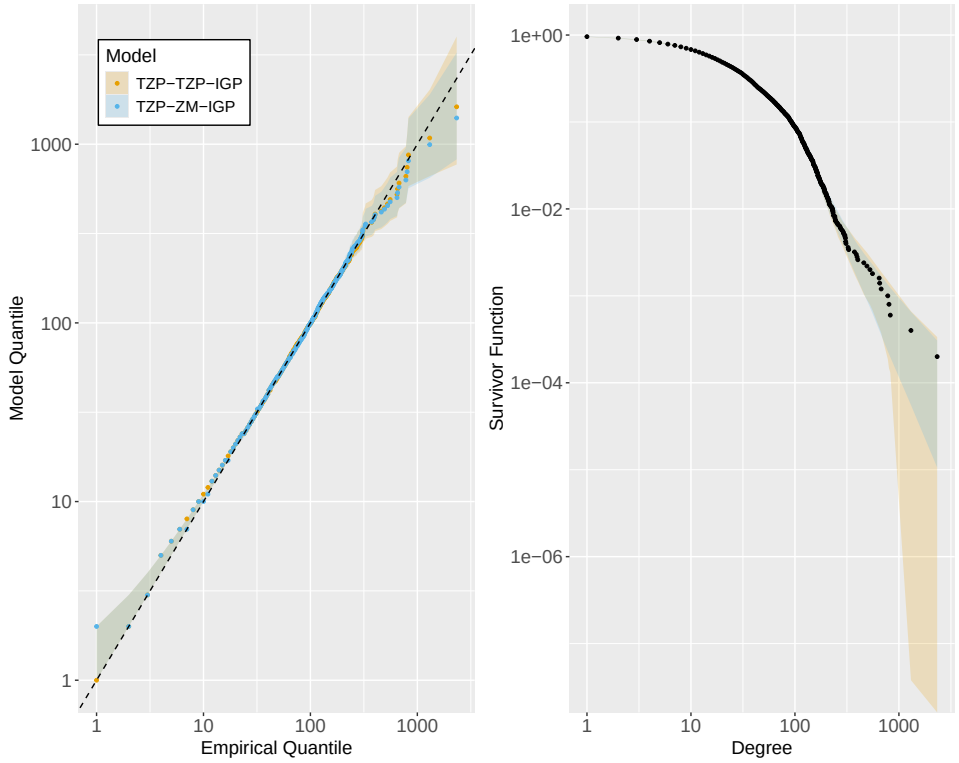

Figure 2.5: QQ-plot (left) and survivor function plot (right), both on the log-log scale, for a single replicate. The 95% confidence intervals, over 200 bootstraps, for the TZP-TZP-IGP and TZP-ZM-IGP are given by the orange and blue bands, respectively. The points in the QQ-plot compare the empirical quantiles and the median model quantiles over the 200 bootstraps. The black dashed line in the QQ-plot represents the $y = x$ line. The black points in the right panel represent the empirical survivor function.

(Bohorquez et al., 2009; Friedman, 2015; Gillespie, 2017), and two sets of word frequencies; "swiss-prot" (Bell et al., 2012) and "moby-dick" (Newman, 2005).

Lee et al. (2024) remove the nodes with degree 1 from the datasets to improve the model fit. Removing them leads to $5 - 60\%$ of the dataset being discarded. On the one hand, there is an argument that nodes with degree 1 may exhibit different behaviour from the remainder of the network, such as bots in online networks or users who make a throwaway account for a one-time offer, and should be treated accordingly. On the other hand, removing them would remove part of the network, which may be informative. Since we aim to model the entire degree distribution, we elect not to remove them.

For each dataset, we fit the two- and three-component mixture models excluding the ZM-TZP-IGP, as this is unlikely to model any of the datasets well in practice. To fit the models, we first perform the threshold selection procedure in Algorithm 2.1 to determine the optimal threshold(s) for each model. Treating the threshold(s) as fixed, the log-likelihoods (2.3.1) and

(2.3.2) are then numerically optimised. Note that the ZM-ZM-IGP was considered, however, it was never chosen as the best-fitting model and has been omitted from further analysis. We also fit the TZP in equation (2.2.1) and determine whether the polylog ($\theta \in (0, 1)$) or the Zipf ($\theta = 1$) distribution is the best-fitting model. For all the datasets we tested, the polylog distribution is deemed the better fit according to AIC and BIC. Comparisons between the polylog and the various mixture models will be made using AIC and BIC (full tables are provided in Supplementary Material) as well as visual diagnostics. Comparisons between the mixture models will also involve comparison of parameter estimates.

Figure 2.6 shows the empirical survivor functions (on the log-log scale) along with 95% confidence intervals, obtained over 200 bootstraps with the thresholds treated as fixed to those previously estimated, for the polylog distribution and selected mixture models. We see that the polylog distribution is inadequate for almost all of the datasets when it is fitted to the entire degree distribution, often resulting in a much lighter tail than suggested by the data. The only dataset the polylog distribution may be appropriate for is "arenas-email", which also minimises the BIC but not the AIC. However, the polylog overestimates the empirical quantiles (see the QQ-plot in Figure 2.7), suggesting the TZP-TZP-IGP is more appropriate.

Although the AIC and BIC agree for most models, they do disagree for "arenas-email", "maayan-vidal", "petster-hamster", "ca-AstroPh", "reactome", "as20000102", "native_american", and "moby". We have already discussed "arenas-email" and how visual diagnostics helped here. We take a similar approach for the other networks to determine the best-fitting model. For instance, for the "maayan-vidal" dataset, the TZP-IGP better captures $y = x$ line in the QQ-plot, and the 95% confidence interval for the survivor function is much tighter compared to the TZP-TZP-IGP. Thus, in the interest of parsimony, the TZP-IGP is deemed the best-fitting model.

Now consider the "moby" dataset. Here, the BIC prefers the ZM-IGP, but the AIC prefers the TZP-ZM-IGP. While we would favour the ZM-IGP in the interest of parsimony, the model drastically overestimates the empirical quantiles, whereas the TZP-ZM-IGP captures them perfectly. Using only the survivor function plot could lead to the ZM-IGP being mistakenly chosen as the best model despite the final few data points being below the lower limit of the 95% confidence interval. This highlights the issue of only assessing model fits using the survivor function plot; small deviations on the log-log scale result in large deviations on the original scale. This may explain why so many networks were inadvertently hailed as "scale-free" following the introduction of the PA model.

As well as considering visual diagnostics, it can be helpful to assess the "penultimate tail"

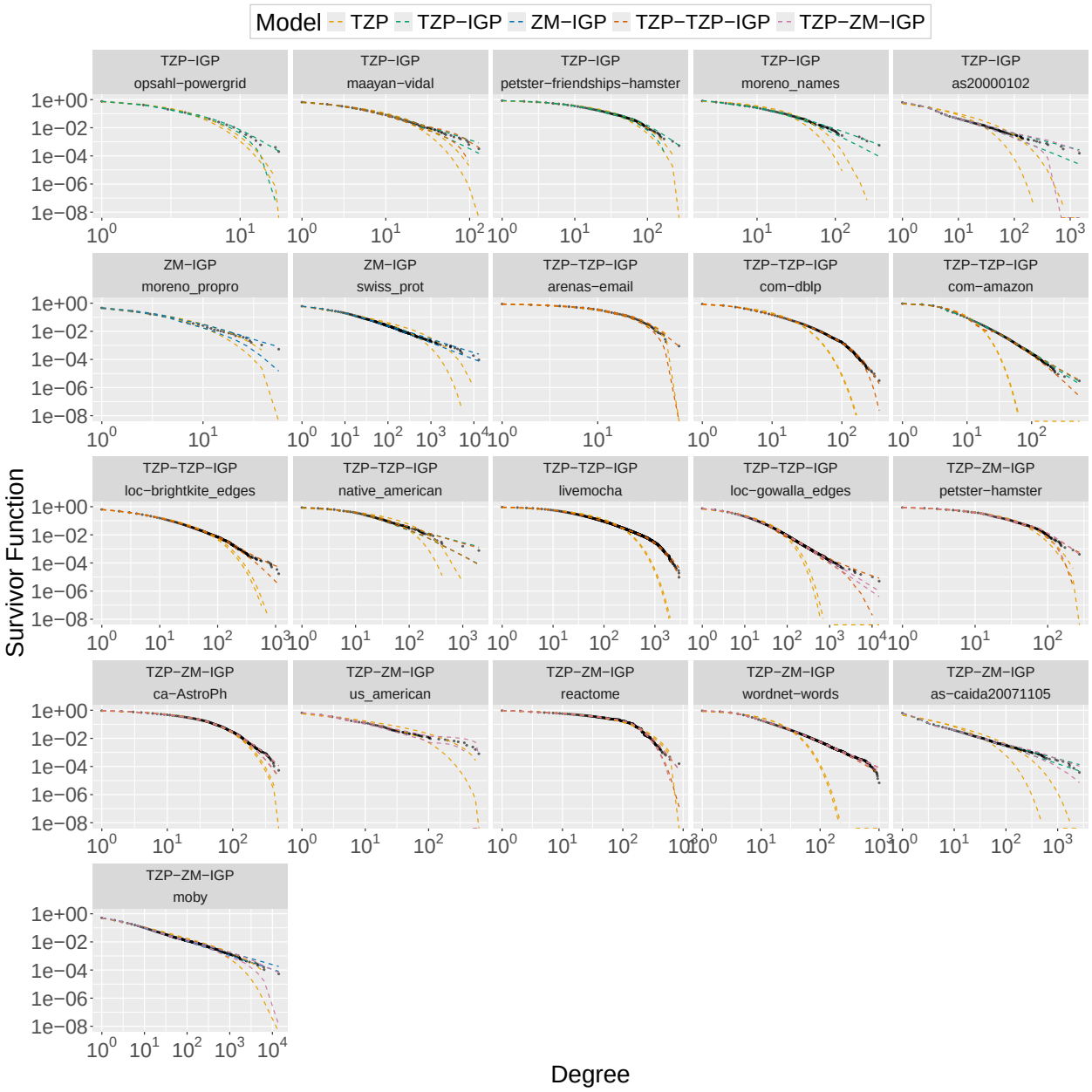

Figure 2.6: Survivor function plots (on the log-log scale) for numerous datasets. The black points represent the empirical survivor function. The dashed lines correspond to 95% confidence intervals, over 200 bootstraps, of the model survivor function. The colour of the lines denotes the model. The best-fitting mixture model is provided above the network name.

(Lee et al., 2024). One can assess the parameter estimates of the component just before the tail to determine whether it is sub-asymptotically power-law or not, or equivalently, if the component is better represented by the ZM or TZP. For example, for the "petster-hamster" dataset $\hat{\theta}_2 = 0.99$ and $\hat{\alpha}_2 = 1.81$ (both to 2dp) for the TZP-TZP-IGP. These MLEs suggest the second component may be modelled better using the ZM distribution.

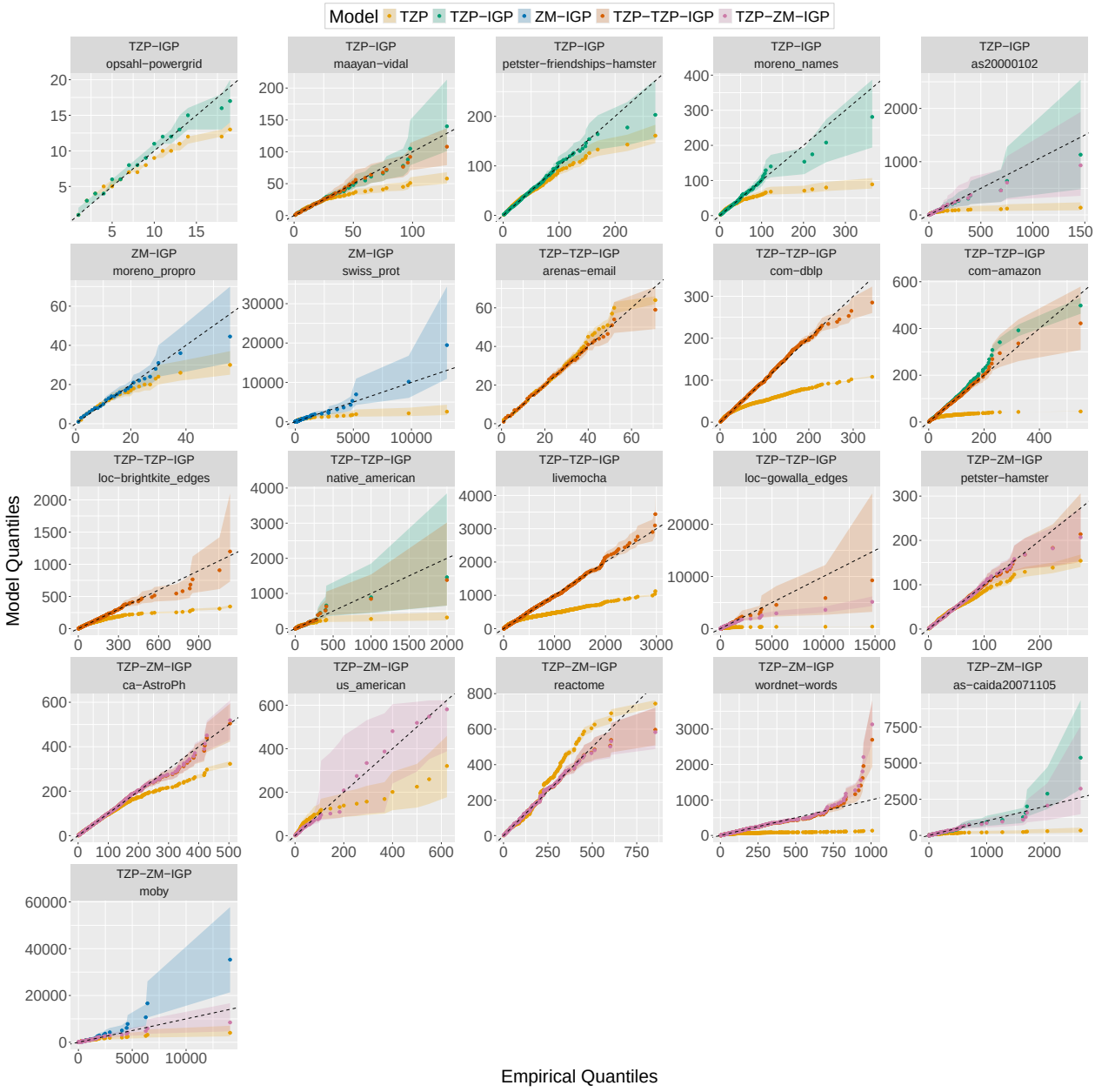

Figure 2.7: QQ-plots for the same data sets in Figure 2.6. The bands in each plot correspond to 95% confidence intervals, over 200 bootstraps, of the model quantiles. The points compare the empirical quantiles and the median model quantile over the 200 bootstrapped samples. The colour of the bands and the points denotes the model. The black dashed lines correspond to the $y = x$ line. The best-fitting mixture model is provided above the network name.

Using the visual diagnostics, one could argue that the TZP-TZP-IGP is the better model, as the 95% confidence interval for the TZP-ZM-IGP fails to cover the $y = x$ line for the most extreme quantiles. However, this is marginal and may be fixed after completing a more comprehensive threshold selection search. In addition, there is little difference in the AIC and BIC (0.6 units for the former and 5.2 units for the latter) for the TZP-TZP-IGP and

TZP-ZM-IGP, suggesting there is little evidence in favour of the more complicated model. Again, in the interest of parsimony, the TZP-ZM-IGP is deemed the best-fitting model in this case. Assessing the other examples where the AIC and BIC disagree similarly, the best mixture model is chosen and provided above the network name in Figures 2.6 and 2.7.

The mixture models appear to fit most of the datasets well except "wordnet-words". Both the TZP-TZP-IGP and TZP-ZM-IGP overestimate the empirical quantiles, particularly in the tail. This is because the threshold selection procedure in Algorithm 2.1 chooses $\hat{v} = 9$ and $\hat{u} = 72$ for the former and $\hat{v} = 8$ and $\hat{u} = 20$ for the latter. This compares with $v \approx 10$ and $u \approx 800$ in Lee et al. (2024). Granted, there are differences in the analysis which may be causing the discrepancy; Lee et al. (2024) removed the 1s from the dataset before fitting, while we retained them. However, the threshold above which the IGP is considered is ultimately too low in both our models and is likely causing bias and a lack of fit. To assess why $\hat{u}$ is so low, Figure 2.8 provides the output from the threshold selection procedure for the TZP-ZM-IGP and TZP-TZP-IGP when applied to the "wordnet-words" dataset. While we expect the bootstrapping procedure to introduce local roughness into the distance metric space, we do not expect the level of roughness presented by the TZP-TZP-IGP. This suggests the model is not appropriate for this dataset. While the distance metric is smooth for the TZP-ZM-IGP, ultimately, the threshold for the tail component is too low. This suggests the space needs to be searched more expansively (we never test $u \approx 800$), or that further components need to be added to better capture the complexity of the data. Other examples where this is also prevalent are "com-dblp", "com-amazon", and "loc-gowalla-edges".

One may question whether the IGP distribution is necessary and whether a simple Zipf distribution may be adequate for the tail component. For several datasets, "opsahl-powergrid", "arenas-email", "us_american", and "com-dblp", the shape parameter of the IGP is negative, suggesting the need for a flexible tail model. The negative shape in these cases is not surprising since the underlying processes should have an upper bound. For instance, the number of casualties cannot exceed the number of troops in "us_american", and the number of power supply lines from a generator, transformer, or substation will be capped for safety purposes in "opsahl-powergrid". Therefore, the IGP is necessary to capture the underlying process behind these datasets.

Alternatively, Lee et al. (2024) compare the implied tail index and the tail index from the IGP to justify the need for a mixture distribution. The implied tail index is interpreted as the tail index had the bulk been extended to encompass the right tail. If the best-fitting mixture model is the TZP-IGP or the TZP-TZP-IGP, then the implied tail index is 0. For the "opsahl-powergrid", "arenas-email", "petster-friendships-hamster", "com-dblp", and

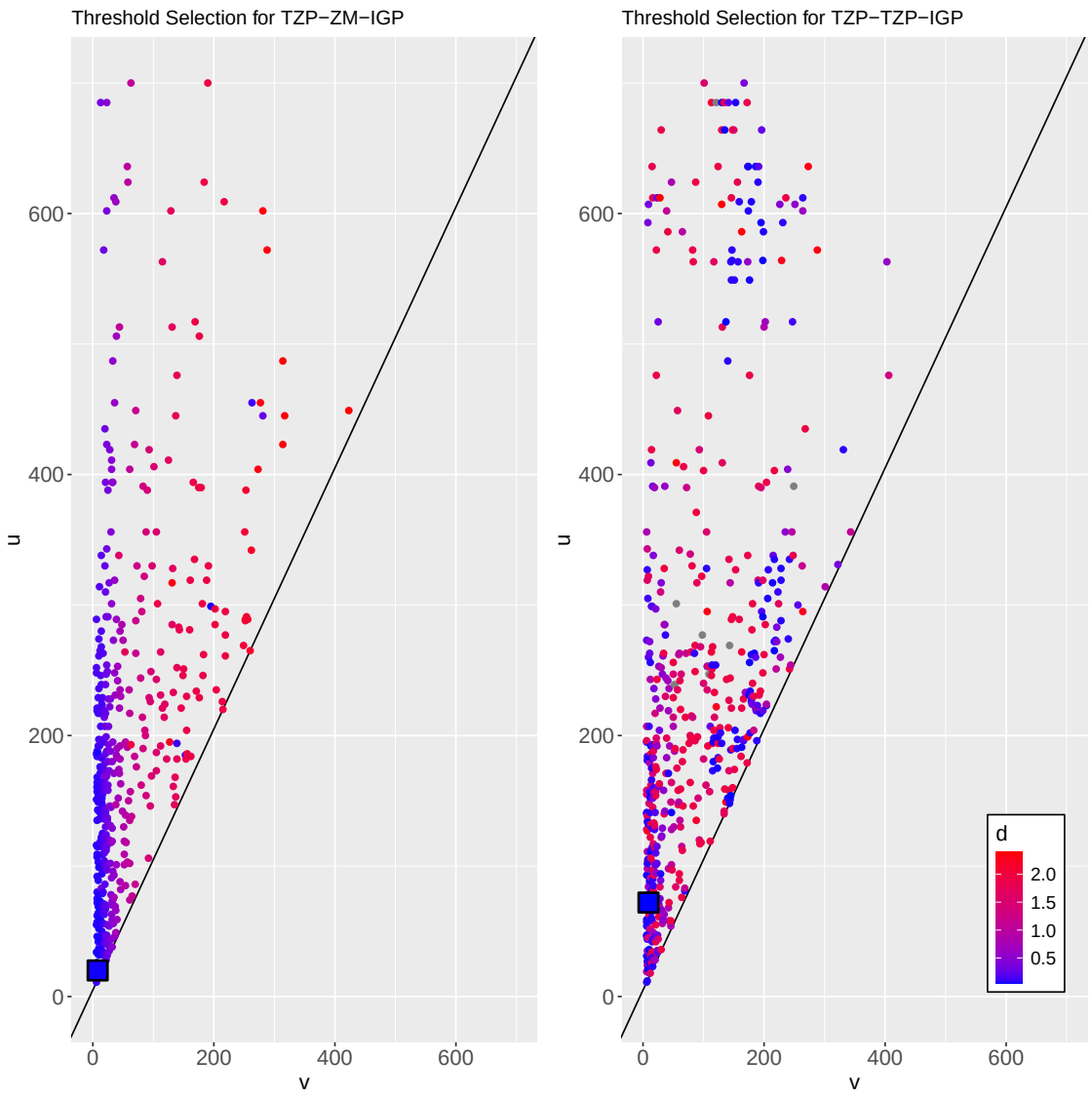

Figure 2.8: Scatter plots of tested candidate thresholds for the TZP-ZM-IGP (left) and TZP-TZP-IGP (right) using the threshold selection procedure in Algorithm 2.1 for the "wordnet-words" dataset. The diagonal black line represents the $u = v + 5$ line (the lower bound for v). The colour of the points details the value of the distance metric. The large square with a black outline is the threshold that minimises the distance metric, amongst those tested, in each case.

"livemocha" datasets, $\hat{\xi} \approx 0$, suggesting the additional flexibility in the tail component may not be necessary. For the remaining datasets where the TZP-IGP or the TZP-TZP-IGP are the best-fitting model, the tail index is much larger than 0, justifying the need for a flexible tail component. If the best-fitting mixture model is the ZM-IGP or the TZP-ZM-IGP, then the implied tail index is $\xi_{\text{mix}} = 1/(\alpha_{\text{mix}} - 1)$, provided $\alpha_{\text{mix}} > 1$ such that $\alpha_{\text{mix}} = \alpha_1$ for the ZM-IGP and $\alpha_{\text{mix}} = \alpha_2$ for the TZP-ZM-IGP. Figure 2.9 compares the tail index and the

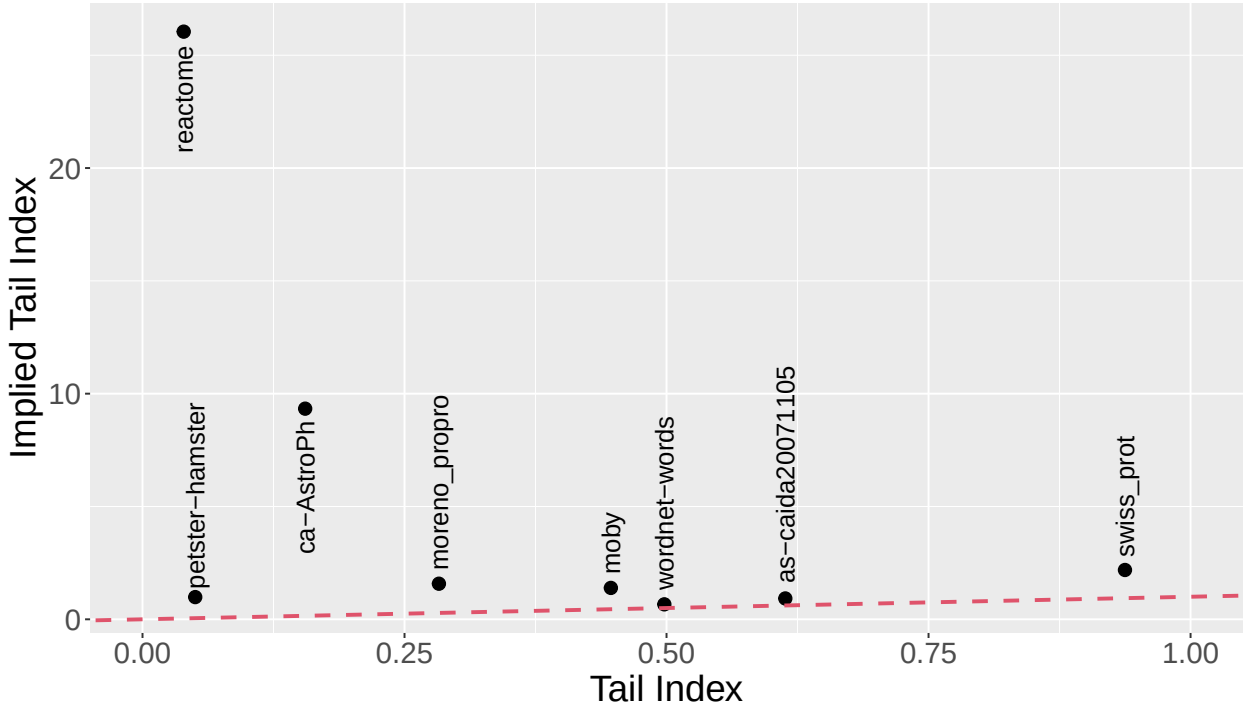

Figure 2.9: Scatter plot comparing the tail index and the implied tail index for selected datasets where the best fitting mixture model is the ZM-IGP or the TZP-ZM-IGP. The network name is provided next to its corresponding point. The red dashed line corresponds to the $y = x$ line.

implied tail index for the datasets where this applies. Note that we have not shown the tail index for the "us_american" dataset as it is very negative ($\hat{\xi} = -1.319$ (3dp)) and has been explained above. For the remaining datasets, the implied tail index is always larger than the tail index, justifying the mixture model to provide a more appropriate rate of tail decay, particularly for the "reactome" and "ca-AstroPh" datasets. Interestingly, the tail index is very similar to the implied tail index for the "wordnet-words" dataset, which further suggests the behaviour of the most extreme degrees is not being accounted for by the mixture model and that the threshold for the tail is too low (see Figure 2.8).

Finally, we note that we attempted to fit larger networks in the KONECT database, such as "com-youtube" and "petster-friendships-dog". However, due to the computational complexity of the proposed threshold selection procedure in Algorithm 2.1, a threshold could not be selected for any of the models. Alternative threshold selection routines are available in the likelihood framework. However, as discussed in Section 2.2.4, they are either more computationally intensive or require a subjective choice of the threshold, and so they are not explored here.

2.6 Discussion

The proposed model appears to fit the data well and is flexible enough to capture a wide range of behaviour. This supports the findings of Lee et al. (2024) and the hypothesis of Mannion and MacCarron (2023) that mixture models are necessary to model the entire degree distribution.

Although the threshold selection method proposed in Section 2.2.4 is computationally more efficient than previous methods, it is still expensive due to two bottlenecks. The first cannot be overcome, as the probability of candidate thresholds being selected for testing changes each iteration depending on which thresholds have previously been tested. The second could be overcome by not refitting the model to the bootstrapped samples in each iteration. An argument against refitting is that if the candidate is a poor choice for the original data, it is likely going to be a poor choice for the bootstrapped samples and result in a high distance metric, irrespective of whether the model is refitted or not. Adopting a Bayesian approach for inference (Lee et al., 2024) is likely to be computationally more efficient since the threshold(s) can be incorporated as parameters to be estimated. In addition, the threshold uncertainty is obtained for free rather than having to repeat the threshold selection procedure in Algorithm 2.1 for many bootstrapped samples.

The Bayesian approach is also more elegant as it allows for model selection via Bayes Factors to be performed within the inference procedure. While a simple likelihood ratio test can be performed to choose the best-fitting mixture model for each candidate threshold, the distance metric used in Algorithm 2.1 is based on absolute deviations in the QQ-plot, meaning it cannot be mixed between different models. Thus, the model selection involves multiple steps and subjective reasoning, making it less statistically robust compared to the Bayesian approach of Lee et al. (2024).

One simplifying assumption, which is unrealistic, is that the degrees of the network are independent. The degree distribution is, by definition, dependent. For example, in a directed network such as Instagram followers, if one user follows an account, this increases the in-degree (followers) of the other account by one, and could increase the in-degree of the original user's account by one if the other user reciprocates. Such dependence should be accounted for when statistically modelling the degree distribution. However, this would require modelling the growth of the network or modelling the degree distribution with a discrete multivariate distribution, both of which are beyond the scope of our contribution and are open questions in the literature. Generative models already account for dependence with directed networks (Bollobás et al., 2003) and notions of reciprocity (Cirkovic et al., 2023).

While proposing generating mechanisms for degree distributions is beyond the scope of our work, it is related. As shown in Section 2.5 and Lee et al. (2024), the mixture models tend to fit the data better than a single distribution, such as Zipf-polylog. This suggests low- and high-degree nodes may have different generating mechanisms that need to be accounted for in the generative models. This raises questions about whether current generative models are appropriate in their current form, as they are designed to capture the limiting behaviour of the largest nodes in the network rather than the network as a whole. In addition, although we can produce networks whose degree distribution decays according to a power-law (Barabási and Albert, 1999) or exponentially (Erdös and Rényi, 1959; Watts and Strogatz, 1998), we are currently unable to generate light-tailed degree distributions. Since certain networks will have an upper bound on their degree distribution, such as the number of junctions in a road and the number of supply lines in power grids, such research is vital to ensure all types of networks can be recreated and understood.

Supplementary Material to "Modelling the Degree Distribution of Networks"

S2.1 Discrete distributions for discrete data

In Section 2.4.1, we showed that the IGP distribution was required for modelling continuous data that had been discretised, and that using the GP distribution resulted in biased parameter estimates. Here, we show that this phenomenon is not isolated to the GP distribution. To do this, we simulate data from the exponential distribution, discretise the data, and model this using the exponential and geometric distributions.

Assume that X follows an exponential distribution with probability density function

$$f(x|\beta) = \begin{cases} \beta\exp(-\beta x) & \text{for } x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

with rate parameter $\beta > 0$. Now, assume that we have a discrete random variable $Y = \lfloor X \rfloor$. It can be shown Y follows a geometric distribution with probability mass function

$$f(y|p) = \begin{cases} p(1-p)^y & \text{for } y \in \{0, 1 \ldots, \}, \\ 0 & \text{otherwise,} \end{cases}$$

where $p = 1 - \exp(-\beta) \in (0, 1]$.

Using a similar approach as in Section 2.4.1, we will simulate 1000 replicates of 1500 realisations from the exponential distribution with $\beta = -\log(0.15) \approx 0.51$ (2dp) such that $p = 0.85$. To obtain the discrete datasets, we will take the integer part of the continuous datasets.

Figure S2.1 compares the parameter estimates obtained from the model fitting process. The left panel compares the MLEs of β when the exponential distribution models both the continuous and discrete datasets. The estimates for the latter case exhibit a large positive bias, while the estimates in the former case are centred around the true value. Similar conclusions can be drawn when assessing the centre panel, which compares estimates when the exponential and geometric distributions model the discrete datasets. The right panel compares MLEs when the exponential and geometric distributions model the continuous and discrete datasets, respectively. Both sets of parameter estimates are centred around the true value (the blue square) and follow the red dashed $y = x$ line. These results are similar to those presented in Section 2.4.1 and bolster the evidence that the only way to obtain unbiased

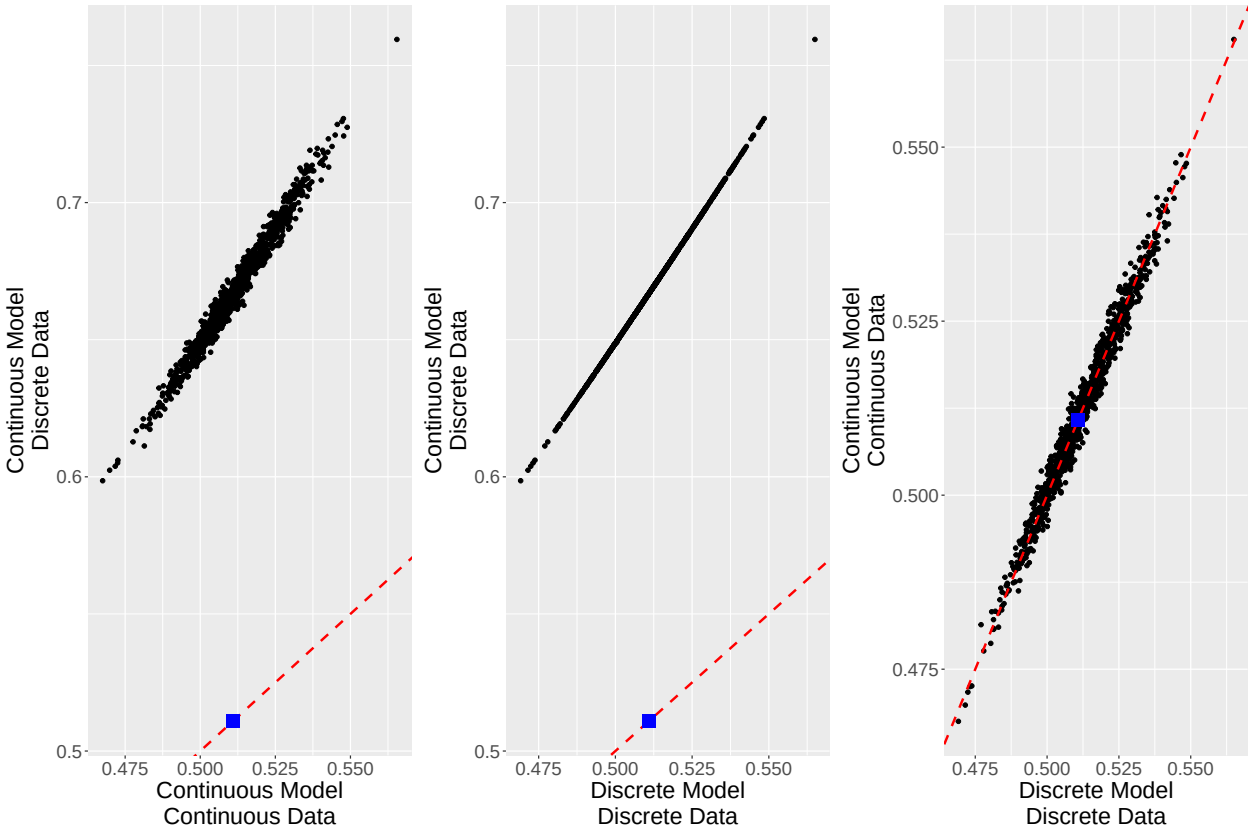

Figure S2.1: Scatter plots of 1000 parameter estimates of β. The left panel compares the MLEs when the exponential distribution models both the continuous and discrete datasets. The centre panel compares the MLEs when the exponential and geometric distributions are fitted to discrete datasets. The right panel compares the MLEs when the exponential distribution models the continuous datasets and the geometric distribution models the discrete datasets. The red dashed lines correspond to the $y = x$ line. The blue square corresponds to the true value of the parameter.

parameter estimates is to use a model that accurately describes the coarseness/fineness of the data.

Similar to Section 2.4.1, Figure S2.2 shows the PP- (left) and QQ-plots (right) for a single randomly selected replicate. Naturally, the PP- and QQ-plots indicate a good model fit when the exponential distribution models the continuous dataset (orange circle) and the geometric distribution models the discrete dataset (green diamonds). When the exponential distribution models the discrete dataset (blue triangles), there is an obvious lack of fit, with the PP-plot exhibiting underestimation, while the QQ-plot exhibits overestimation for low quantiles and underestimation for high quantiles.

Therefore, erroneously using a continuous distribution to model discrete data can result in a deceptively good-fitting model, as shown in Section 2.4.1. However, it can also result in

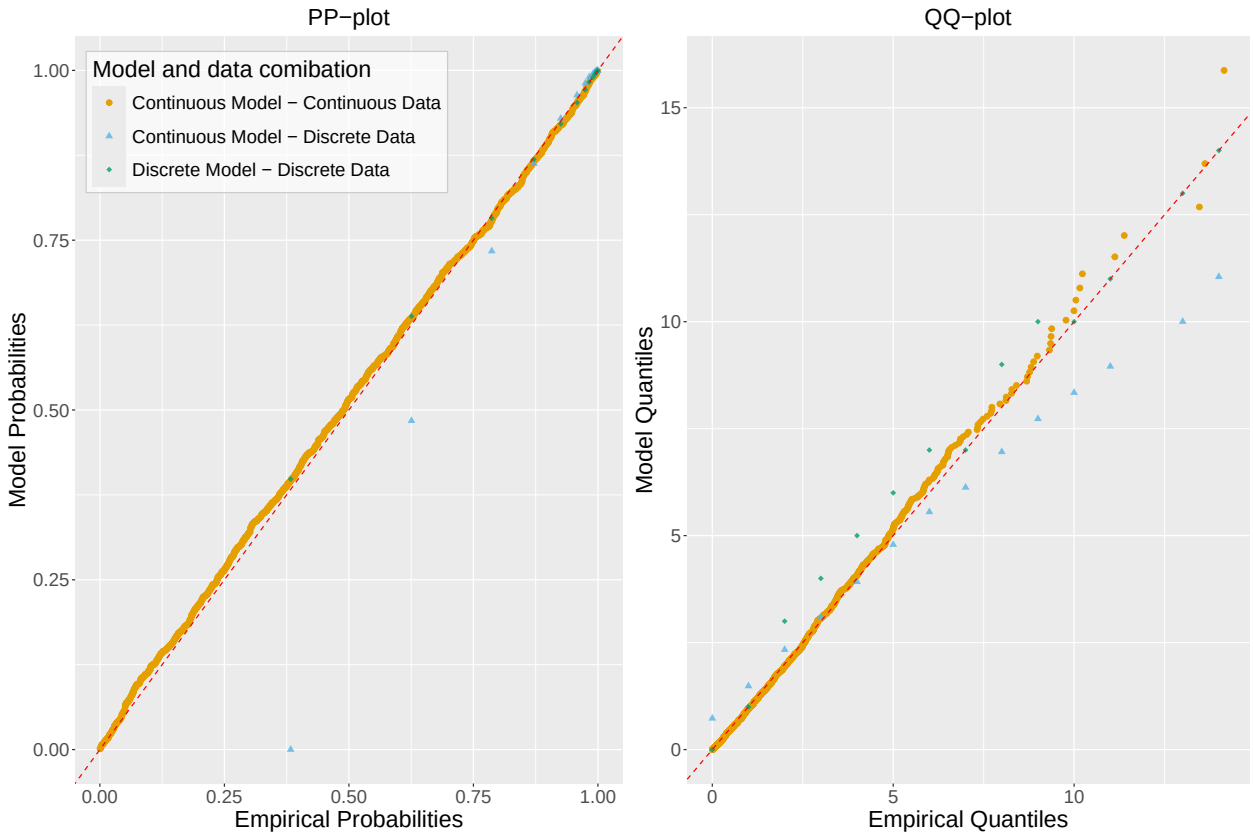

Figure S2.2: PP- (left) and QQ-plot (right) for a single replicate when the exponential distribution is used to model the continuous (orange circles) and discrete (blue triangles) datasets, and when the geometric distribution is used to model the discrete datasets (green diamonds). The red dashed line represents the $y = x$ line.

an ill-fitting model, as shown in Figure S2.2. The alignment of the PP- and QQ-plot with the $y = x$ line will broadly depend on the range of the data being considered and whether the discrete random variable can be assumed to be "approximately" continuous. Given the range of discrete distributions available and the ease with which they can be fit, such false assumptions should not need to be made in practice.

S2.2 Approximating discrete random variables

In the network science literature, it is common practice to model discrete data using a continuous distribution after the data has been perturbed by some random noise (Voitalov et al., 2019). Here, we show that this does not result in biased parameter estimates, however, the correct assumptions must be made.

Assume that an integer-valued random variable Y follows the IGP under ceiling above some

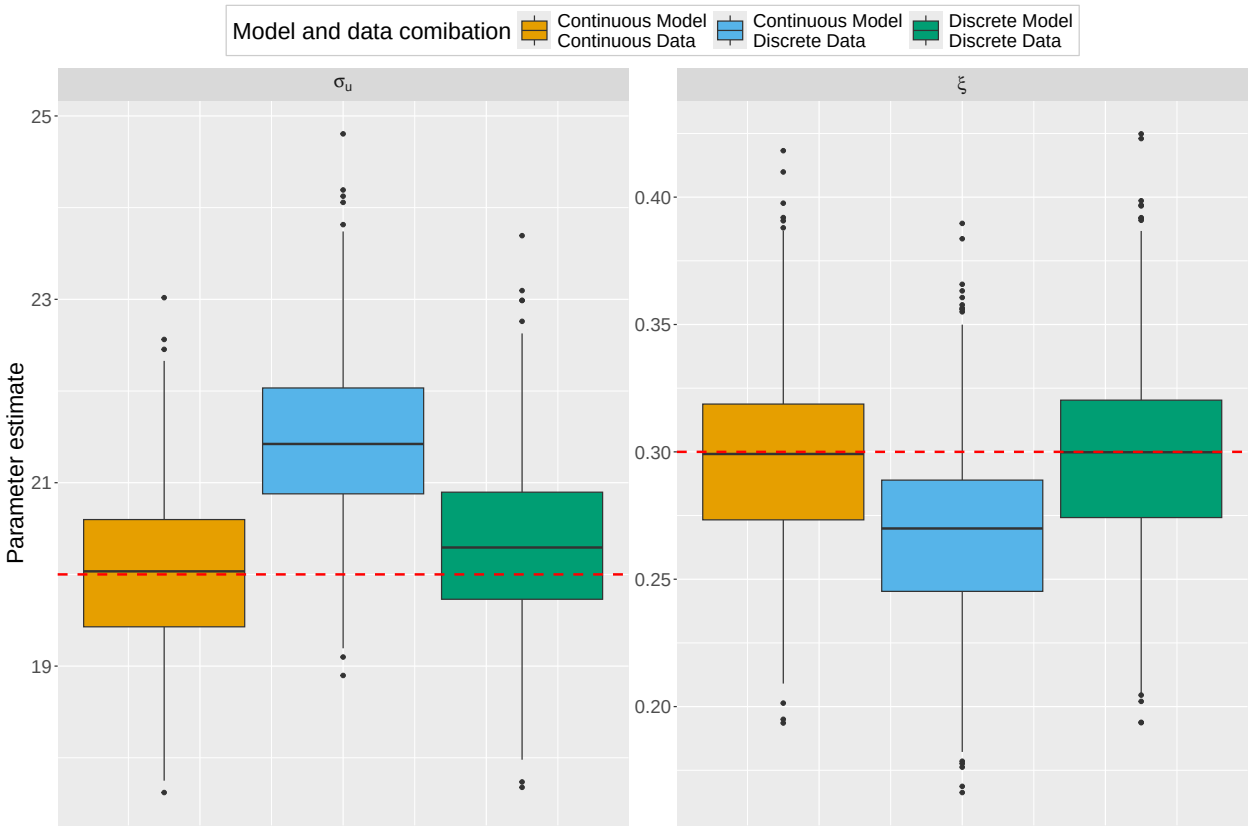

Figure S2.3: Boxplots of 1000 parameter estimates (scale - left, shape - right) from when the GP distribution models continuous (orange) and discrete (blue) datasets and when the IGP under flooring models discrete datasets (green). The red dashed lines correspond to the true parameter values.

high threshold u. To make the data continuous, we need to perturb the data. We cannot add uniform noise, since we assume Y follows the IGP under ceiling, however, we cannot subtract uniform noise either, as $X = Y - U$, where U is a uniform random variable on the interval $(0, 1)$, would follow a continuous GP above a threshold $u - 1$. Hence, X and Y are not consistent. Therefore, we have to assume that Y is an integer-valued random variable that follows the IGP under flooring above some high threshold u. Again, by definition, we cannot assume that $X = Y - U$. Thus, we ***must*** assume that $X = Y + U$.

Assume that Y is an integer-valued random variable that follows the IGP under *flooring* above some high threshold $u = 10$ with scale parameter $\sigma_u = 20$, and shape parameter $\xi = 0.3$. As per Section 2.4.1, we generate 1000 replicates of $n = 1500$ realisation from the model, such that $y_i^{(j)}$ is the ith realisation from the jth replicate, for $i = 1, \ldots, 1500$ and $j = 1000$. For the continuous data, define $x_i^{(j)} = y_i^{(j)} + u_i^{(j)}$, where $u_i^{(j)}$ is a realisation from a standard uniform distribution. We can therefore assume that X follows a GP distribution with threshold u, scale parameter σ_u and shape parameter ξ.

Figure S2.3 shows boxplots of the parameter estimates when the GP distribution models the continuous (orange) and the discrete datasets (blue), and when the IGP under flooring models the discrete datasets (green). The second combination is designed to provide biased parameter estimates, since the model assumes the data is on a finer scale than is available. The first and third combinations are unbiased, and this is the ***only*** way to obtain unbiased parameter estimates when modelling discrete extreme values using a GP distribution.

S2.3 Additional threshold selection examples

In Section 2.4.2, we assessed the proposed threshold selection method in Algorithm 2.1 for the TZP-ZM-IGP distribution. In Section S2.3.1, we repeat the process here for the ZM-IGP distribution to evaluate the performance in an edge case setting. In addition, we perform the threshold selection for data generated from the geometric distribution in Section S2.3.2 to assess how our model performs for data not generated from the underlying mixture distribution.

S2.3.1 ZM-IGP threshold selection

Similar to Section 2.4.2, we simulate 100 replicates of size $50,000$ from the ZM-IGP with threshold $u = 35$ (alternatively $\phi_u = 0.11$), and parameters $\theta = 1$, $\alpha = 1.5$, $\sigma_u = 55$, and $\xi = 0.8$. Such parameters are chosen to make it difficult to determine where the body ends and the tail begins.

For each simulated dataset, the threshold selection method in Section 2.2.4 is applied for both the ZM-IGP and TZP-IGP. Note, in Algorithm 2.1, the number of bootstraps B is set to 200, the number of tested thresholds at each iteration k is set to 10, and the number of consecutive trials without reducing the distance metric s is also set to 10. This results in approximately 25% of candidate thresholds being tested, on average, thereby providing a large computational gain compared to testing all possible candidate thresholds. The thresholds are then treated as fixed, and the model is fitted using the log-likelihood (2.3.1).

Figure S2.4 shows boxplots of the threshold and parameter estimates for the ZM-IGP (orange) and TZP-IGP (blue). For the TZP-IGP, $\theta \approx 1$ and $\alpha > 1$ for all replicates. This indicates that the ZM-IGP may be a more appropriate model. Both models overestimate the true threshold, although the ZM-IGP is less biased than the TZP-IGP. Overestimation in the threshold results in overestimation in the scale parameter of the IGP, since it is threshold-dependent, and a slight underestimation of α in the bulk. Since the selected threshold exceeds the true threshold, the shape parameter is unbiased for both models due to the parameter

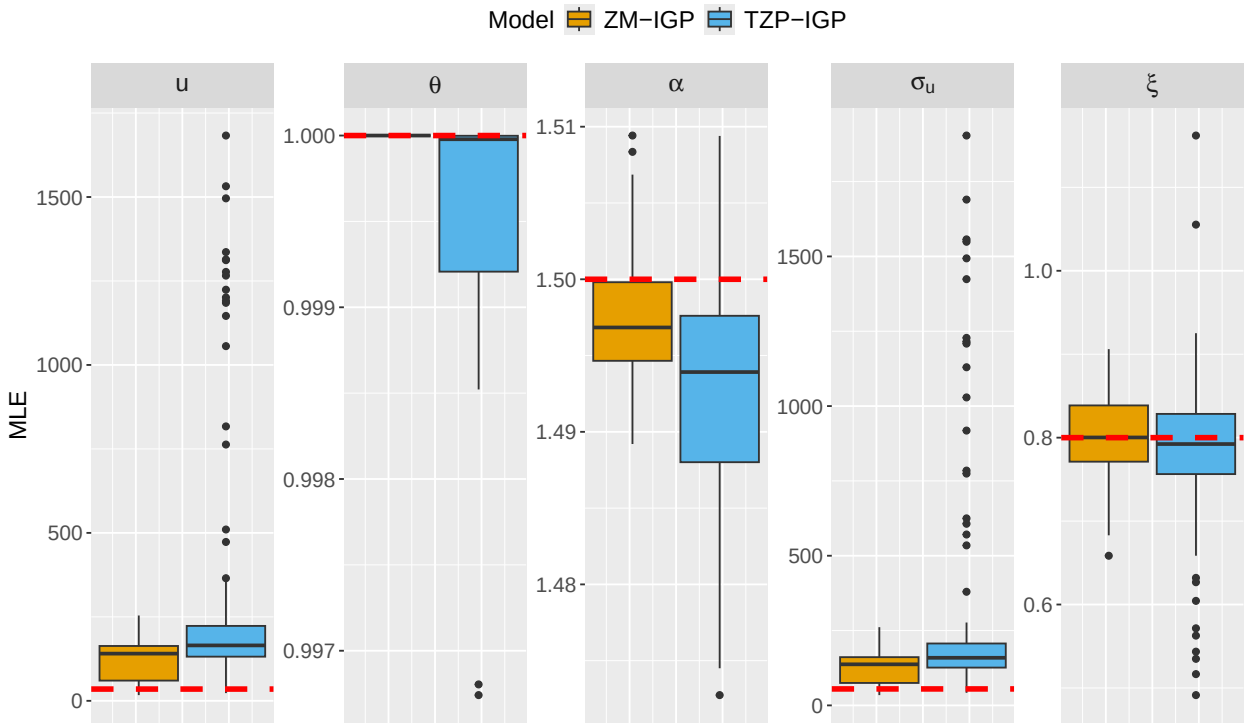

Figure S2.4: Boxplots of the selected threshold and subsequent parameter estimates for the ZM-IGP (orange) and TZP-IGP (blue). The threshold and parameters are u (left), θ (centre left), α (centre), σ_u (centre right), and ξ (right). The red dashed lines in each panel correspond to the true threshold/parameter value.

stability property of the IGP.

Figure S2.5 shows the degree distribution, survivor function, and distance metric output (all on the log-log scale) for a randomly selected replicate. From the first two plots, the distinction of where the body ends and the tail begins is unclear, which explains why the threshold selection method overestimates the true threshold. The distance metric for the ZM-IGP demonstrates the shape we would expect; a decreasing distance metric below the true threshold due to bias in the IGP component, an increasing distance metric above the selected threshold due to increased variability in the IGP component and bias in the ZM component, and a relatively flat metric between the two thresholds where any of the candidate thresholds could be suitable (although there is an odd bump around 100). The TZP-IGP somewhat reflects this shape, but the distance metric has an inexplicable bump around 250. The roughness in the curve is quite clear and highlights why gradient methods are not appropriate for minimising the distance metric. The curve could be smoothed further by increasing the number of bootstraps; however, this increases the computational cost without drastically improving the result.

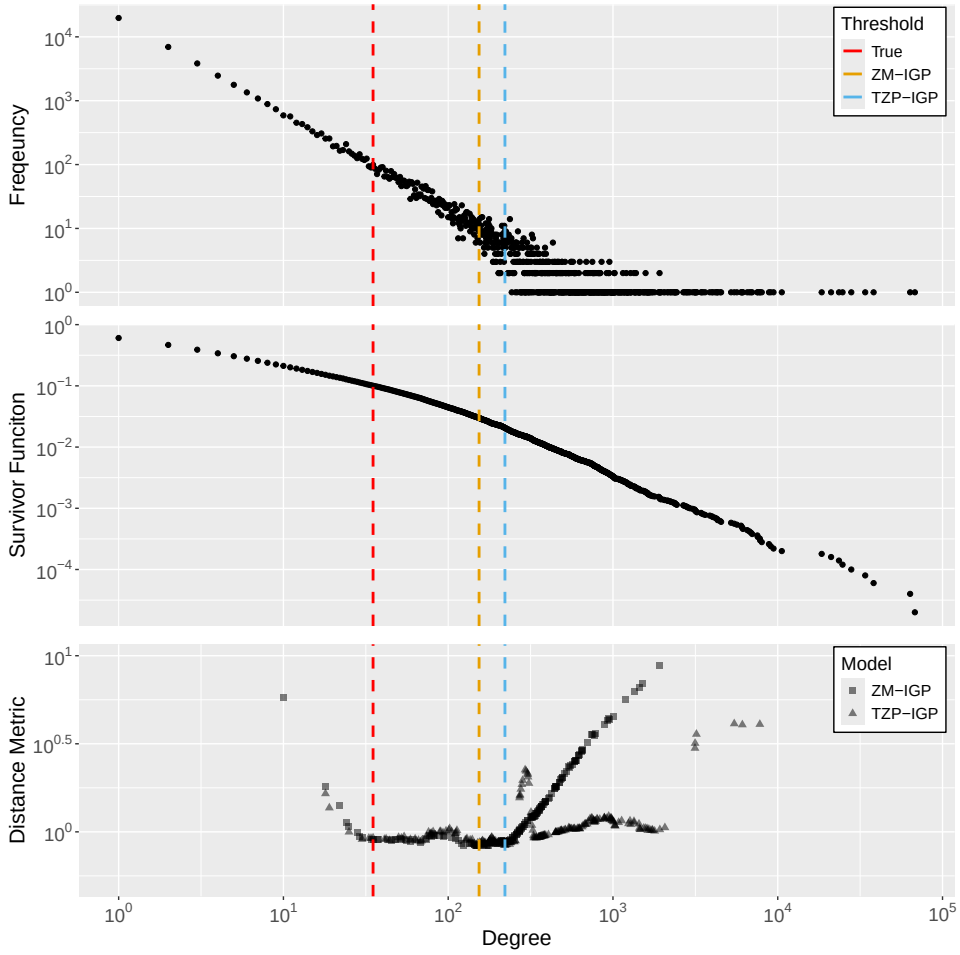

Figure S2.5: The degree distribution (top), survivor function (middle), and the distance metric (bottom) on the log-log scale, for a single replicate. The red, orange, and blue dashed vertical lines correspond to the true threshold and the selected thresholds for the ZM-IGP and TZP-IGP, respectively. The squares and triangles in the bottom panel correspond to the distance metric for the ZM-IGP and TZP-IGP, respectively.

Despite both models overestimating the true threshold, the model fit is not poor, as can be seen by the QQ-plot in Figure S2.6. The QQ-plot is on the log-log scale to highlight both tails. Both models overestimate the left tail, which is likely due to the α parameter being slightly underestimated. The right tail has some deviations, although nothing unexpected, given the heavy nature of the tail in question. Despite being simulated from the ZM-IGP, the fit would likely improve by using the three-component mixture model.

Finally, we would need to choose the model that best fits the data. Given that the model fits are visually comparable, we need to use alternative methods. Comparing the models using AIC and BIC, they correctly choose the ZM-IGP as the best-fitting model 87 and 94 times, respectively. In addition, $\hat{\theta} \approx 1$ and $\hat{\alpha} > 1$ for all replicates (see Figure S2.4), which gives a

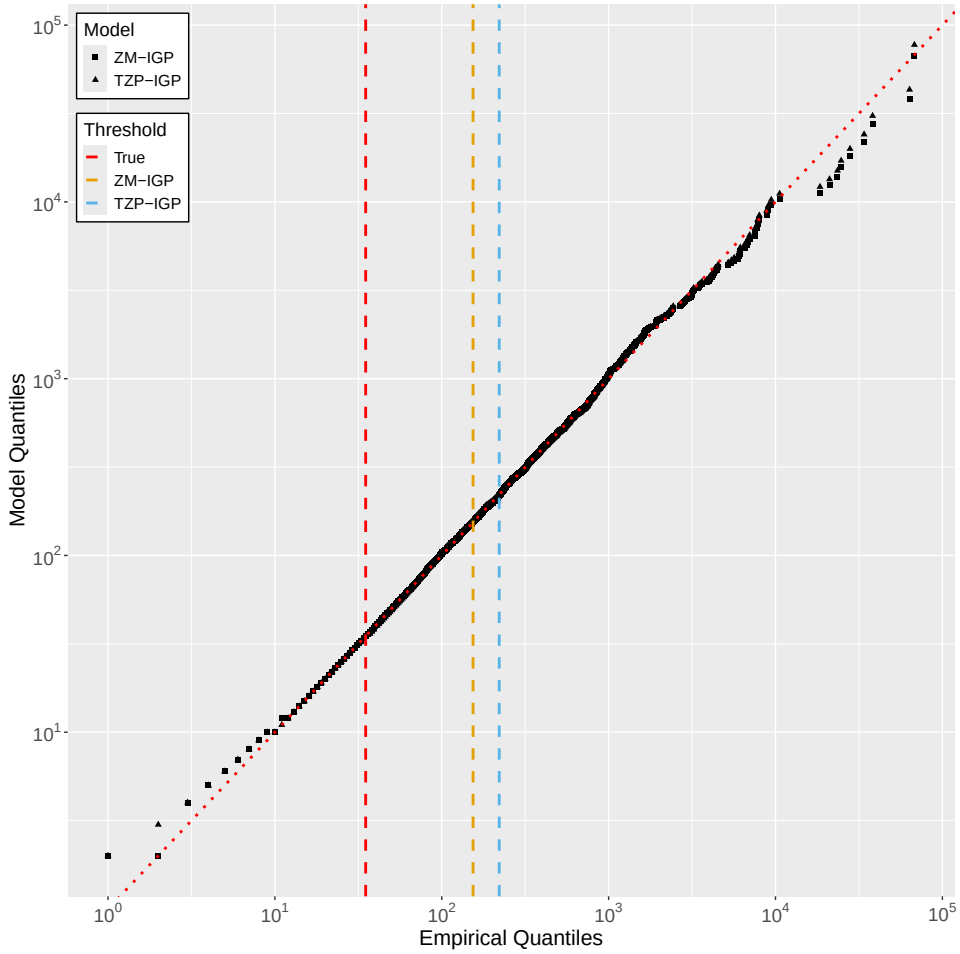

Figure S2.6: QQ-plot (on the log-log scale) comparing the model fit from the ZM-IGP (squares) and TZP-IGP (triangles) for a single replicate. The red, orange, and blue dashed vertical lines correspond to the true threshold and the selected thresholds for the ZM-IGP and TZP-IGP, respectively. The red dotted line corresponds to the $y = x$ line.

strong indication that the ZM-IGP is the more appropriate model.

S2.3.2 Geometric threshold selection

For this simulation study, we simulated 100 replicates of size 5000 from the geometric distribution with the probability of success $p = 0.05$. For each simulated dataset, we perform the threshold selection method in Section 2.2.4 for the TZP-IGP only. The ZM-IGP is inappropriate for data generated from a geometric distribution, as the bulk will not exhibit power-law-like behaviour. Further, more complicated mixtures, such as the TZP-TZP-IGP, are likely to overfit to the data. Note, in Algorithm 2.1, the number of bootstraps B is set to 500, the number of tested thresholds at each iteration k is set to 5, and the number of consecutive trials without reducing the distance metric s is also set to 5. This results in ap-

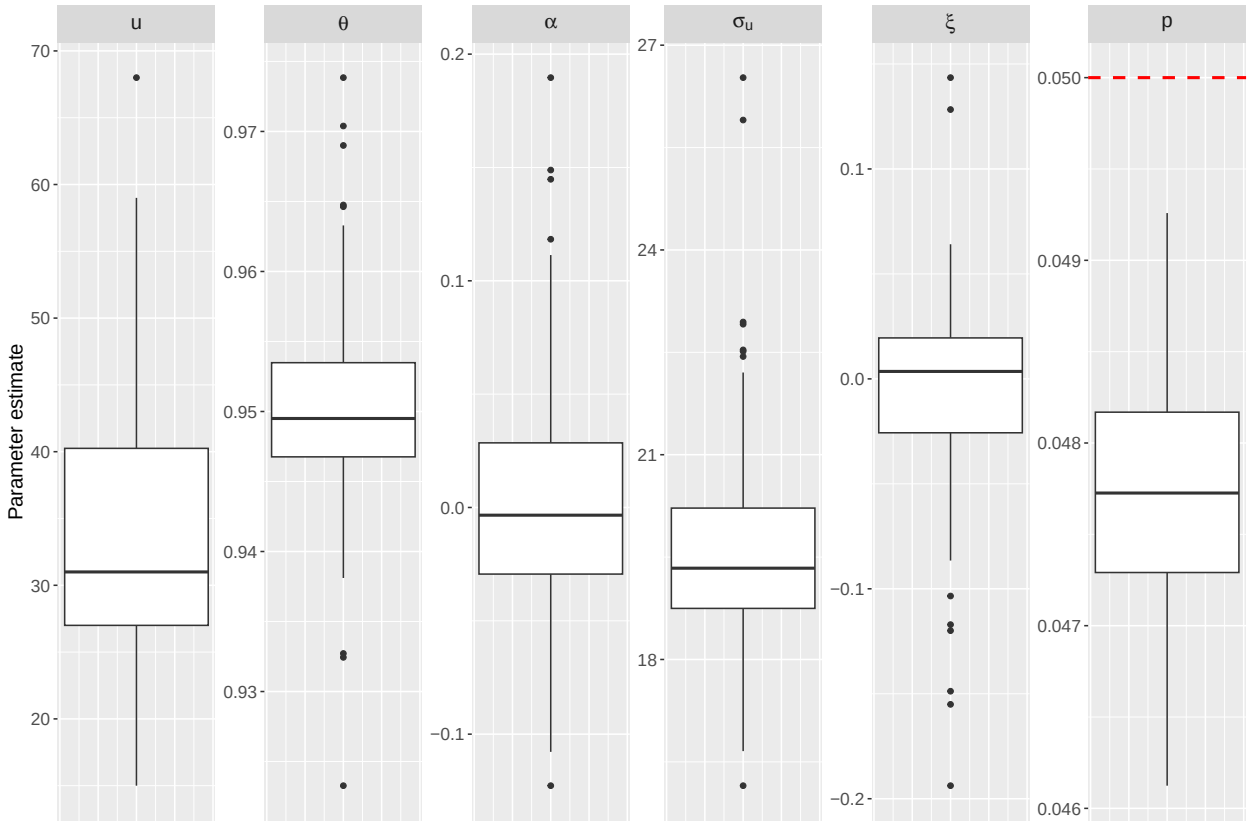

Figure S2.7: Boxplots of the selected threshold and subsequent parameter estimates for the TZP-IGP and the parameter estimate for the geometric distribution. The threshold and parameters are u (first), θ (second), α (third), σ_u (fourth), ξ (fifth), and p (sixth). The red dashed line in the final panel corresponds to the true parameter value.

proximately 45% of candidate thresholds being tested, on average, thereby providing a large computational gain compared to testing all possible candidate thresholds. The thresholds are then treated as fixed, and the model is fitted using the log-likelihood (2.3.1). For comparison, we also fit the log-likelihood function for the geometric distribution.

The selected threshold and parameter estimates from the TZP-IGP, as well as the parameter estimate for p in the geometric distribution, are provided in Figure S2.7. First, note that the probability of success p for the geometric distribution is underestimated compared to its true value. This is likely due to the small value of p and the large range of the data being considered. However, the bias is small enough that it does not cause concern. For the selected threshold u, there is quite a range of values, which may be due to the smaller sample size and the small value of p. For the truncated polylog component, the θ parameter is sufficiently far away from 1 that we do not need to consider the ZM-IGP as an alternative model. For the IGP component, the shaper parameter is 0 on average. While the geometric distribution does not belong to the domain of attraction of any distribution (Shimura, 2012),

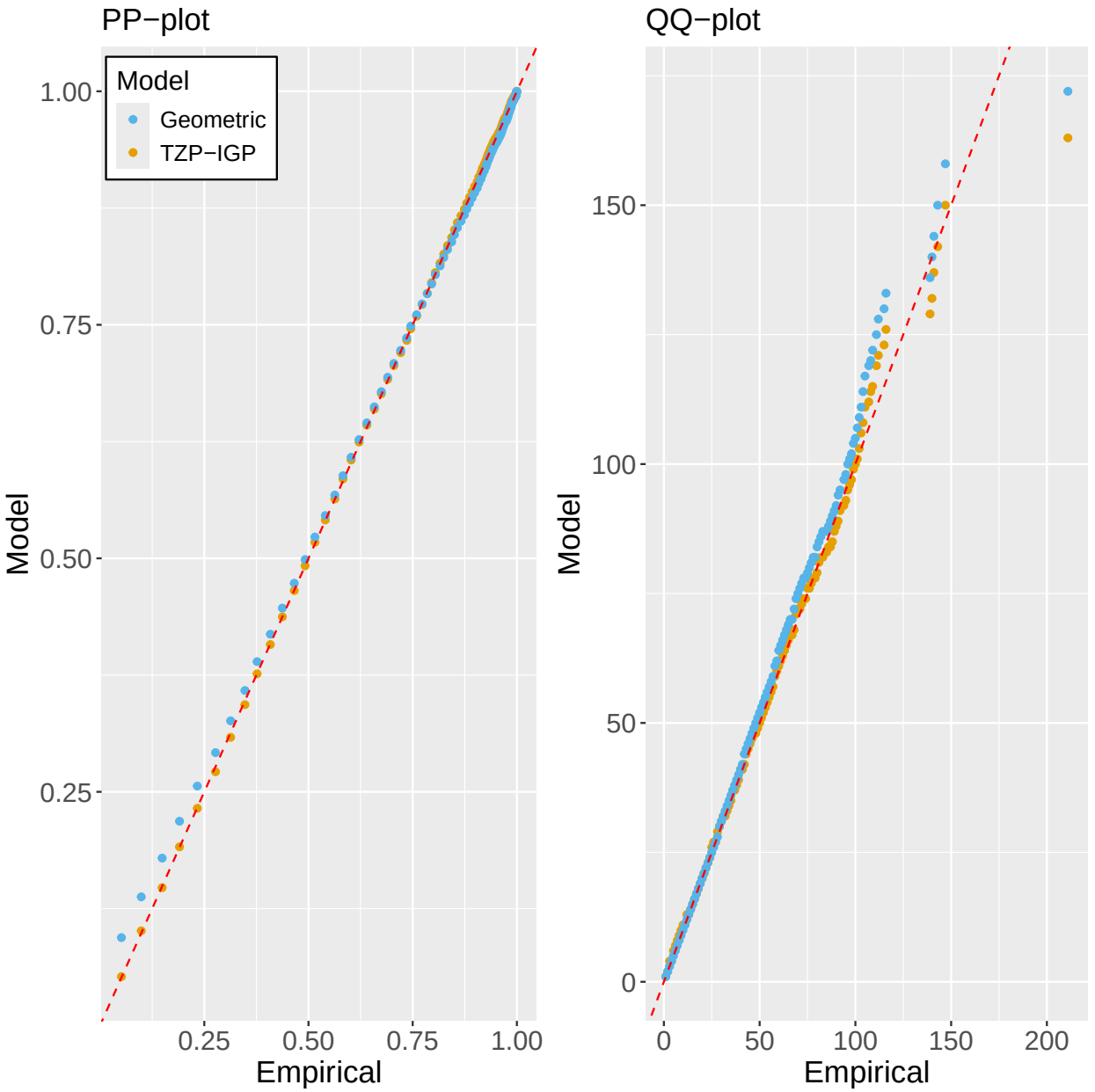

Figure S2.8: PP- (left) and QQ-plot (right) comparing the model fit from the geometric (blue) and TZP-IGP (orange) distributions for a single replicate from the geometric distribution. The red dashed line corresponds to the $y = x$ line.

a shape parameter close to 0 makes sense since the geometric distribution is the discrete counterpart of the exponential distribution, which belongs to the domain of attraction of the Gumbel distribution.

The parameter estimates appear sensible, however, they do not inform us of the fit of the models. Figure S2.8 shows the PP- and QQ-plots for a single replicate for both the geometric and TZP-IGP distributions. Interestingly, the low empirical probabilities are overestimated

by the geometric distribution, a consequence of the slightly biased parameter estimate. Furthermore, the largest empirical probabilities appear to be underestimated, which is clearer on the QQ-plot, where the largest quantiles are overestimated. Conversely, the PP-plot for the TZP-IGP presents no issues, and the overestimation of the largest quantiles for this model is a little lower. This suggests that the mixture model provides a good fit, despite the data itself not being generated from a mixture distribution. This also indicates the threshold selection is performing adequately to provide a better fit to both the bulk and the tail compared to using a single distribution for both. This is supported by the AIC and BIC choosing the TZP-IGP over the geometric distribution for all 100 replicates despite having three additional parameters. Thus, the threshold selection appears be choosing a sensible threshold and optimising the overall fit to the data.

S2.4 AIC and BIC tables for the application

This section contains tables to supplement the model selection performed in Section 2.5. Table S2.1 contains the AIC for the TZP, and the change in AIC, relative to the TZP model, for all the other models. Table S2.2 details the same information but for the BIC. The bold number in each row denotes the model that minimises the AIC/BIC.

Table S2.1: The first column details the AIC for the TZP for various datasets analysed in Section 2.5. The remaining columns depict the change in AIC relative to the TZP model. The bold number in each row corresponds to the model that minimises the AIC.

Network Name	TZP	ZM	TZP-IGP	ZM-IGP	TZP-TZP-IGP	TZP-ZM-IGP
opsahl-powergrid	1.70×10^4	2.46×10^3	$\mathbf{-169}$	784	NA	NA
moreno_propro	5.84×10^3	96.8	-9.08	$\mathbf{-9.38}$	-8.98	NA
arenas-email	7.27×10^3	346	-4.76	40.5	$\mathbf{-4.94}$	NA
maayan-vidal	1.40×10^4	465	-94.4	-71	$\mathbf{-95.3}$	NA
petster-friendships-hamster	1.29×10^4	537	$\mathbf{-41.5}$	67.4	-40.7	NA
petster-hamster	1.70×10^4	870	-54.7	75.7	$\mathbf{-64.2}$	-63.6
com-dblp	1.77×10^6	1.35×10^5	-4.38×10^4	9.47×10^3	$\mathbf{-4.49 \times 10^4}$	-4.48×10^4
moreno_names	1.16×10^4	993	$\mathbf{-673}$	218	-639	-641
ca-AstroPh	1.48×10^5	7.93×10^3	-787	1.58×10^3	$\mathbf{-964}$	-961
com-amazon	1.72×10^6	3.46×10^5	$\mathbf{-7.27 \times 10^4}$	1.07×10^5	-3.52×10^4	-6.12×10^4
us_american	6.05×10^3	23.5	-60.8	-89.8	-94.5	$\mathbf{-110}$
reactome	5.84×10^4	1.80×10^3	-284	104	$\mathbf{-640}$	-636
wordnet-words	9.10×10^5	8.29×10^4	-4.12×10^4	-5550	-4.13×10^4	$\mathbf{-4.13 \times 10^4}$
loc-brightkite_edges	2.96×10^5	4700	-638	-548	$\mathbf{-671}$	-592
as20000102	2.64×10^4	337	-1300	-1140	-1300	$\mathbf{-1300}$
native_american	9.52×10^3	361	-180	-109	$\mathbf{-185}$	-182
as-caida20071105	9.59×10^4	366	$\mathbf{-9000}$	-3310	-8490	-8490
livemocha	9.28×10^5	4.16×10^4	-1.36×10^4	2830	$\mathbf{-1.45 \times 10^4}$	-1.42×10^4
swiss_prot	5.86×10^4	180	-176	$\mathbf{-187}$	-174	-177
moby	8.03×10^4	33.8	-174	-171	-174	$\mathbf{-179}$
loc-gowalla_edges	1.15×10^6	3.68×10^4	-1.20×10^4	-1.04×10^4	-8090	$\mathbf{-1.22 \times 10^4}$

Table S2.2: The first column details the BIC for the TZP for various datasets analysed in Section 2.5. The remaining columns depict the change in BIC relative to the TZP model. The bold number in each row corresponds to the model that minimises the BIC.

Network Name	TZP	ZM	TZP-IGP	ZM-IGP	TZP-TZP-IGP	TZP-ZM-IGP
opsahl-powergrid	1.71×10^4	2.45×10^3	$\mathbf{-156}$	790	NA	NA
moreno_propro	5.85×10^3	91.3	1.99	$\mathbf{-3.85}$	13.2	NA
arenas-email	$\mathbf{7.28 \times 10^3}$	341	5.31	45.5	15.2	NA
maayan-vidal	1.40×10^4	459	$\mathbf{-82.3}$	-65	-71.1	NA
petster-friendships-hamster	1.29×10^4	532	$\mathbf{-30.5}$	72.9	-18.6	NA
petster-hamster	1.71×10^4	865	-43.2	81.5	-41	$\mathbf{-46.2}$
com-dblp	1.77×10^6	1.35×10^5	-4.37×10^4	9.48×10^3	$\mathbf{-4.48 \times 10^4}$	-4.48×10^4
moreno_names	1.16×10^4	987	$\mathbf{-662}$	224	-617	-624
ca-AstroPh	1.48×10^5	7.92×10^3	-771	1.59×10^3	-932	$\mathbf{-937}$
com-amazon	1.72×10^6	3.46×10^5	$\mathbf{-7.26 \times 10^4}$	1.07×10^5	-3.51×10^4	-6.12×10^4
us_american	6.06×10^3	18.4	-50.6	-84.6	-74	$\mathbf{-94.3}$
reactome	5.84×10^4	1.79×10^3	-271	111	-613	$\mathbf{-616}$
wordnet-words	9.10×10^5	8.29×10^4	-4.12×10^4	-5540	-4.12×10^4	$\mathbf{-4.13 \times 10^4}$
loc-brightkite_edges	2.96×10^5	4690	-620	-539	$\mathbf{-635}$	-565
as20000102	2.64×10^4	330	$\mathbf{-1280}$	-1140	-1270	-1280
native_american	9.53×10^3	356	$\mathbf{-169}$	-104	-164	-167
as-caida20071105	9.59×10^4	358	$\mathbf{-8980}$	-3300	-8460	-8470
livemocha	9.28×10^5	4.15×10^4	-1.36×10^4	2830	$\mathbf{-1.45 \times 10^4}$	-1.42×10^4
swiss_prot	5.86×10^4	173	-161	$\mathbf{-180}$	-145	-155
moby	8.04×10^4	25.9	-158	$\mathbf{-163}$	-142	-155
loc-gowalla_edges	1.15×10^6	3.68×10^4	-1.20×10^4	-1.04×10^4	-8050	$\mathbf{-1.21 \times 10^4}$

Bibliography

André, L., Wadsworth, J., and O'Hagan, A. (2024). Joint modelling of the body and tail of bivariate data. *Computational Statistics & Data Analysis*, 189:107841.

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244.

Bell, M. J., Gillespie, C. S., Swan, D., and Lord, P. (2012). An approach to describing and analysing bulk biological annotation quality: a case study using UniProtKB. *Bioinformatics*, 28(18):i562–i568.

Bohorquez, J. C., Gourley, S., Dixon, A. R., Spagat, M., and Johnson, N. F. (2009). Common ecology quantifies human insurgency. *Nature*, 462(7275):911–914.

Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290.

Bollobás, B., Borgs, C., Chayes, J., and Riordan, O. (2003). Directed scale-free graphs. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 132–139, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, 33(1):309–320.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, 10(1).

Carreau, J. and Bengio, Y. (2009). A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12(1):53–76.

Castro-Camilo, D., Huser, R., and Rue, H. (2019). A spliced gamma-generalized Pareto model for short-term extreme wind speed probabilistic forecasting. *Journal of Agricultural, Biological, and Environmental Statistics*, 24(3):pp. 517–534.

Chattopadhyay, S., Chakraborty, T., Ghosh, K., and Das, A. K. (2021). Uncovering patterns in heavy-tailed networks: A journey beyond scale-free. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, CODS-COMAD '21, page 136–144, New York, NY, USA. Association for Computing Machinery.

Chattopadhyay, S., Murthy, C., and Pal, S. K. (2014). Fitting truncated geometric distributions in large scale real world networks. *Theoretical Computer Science*, 551:22–38.

Cirkovic, D., Wang, T., and Resnick, S. I. (2023). Preferential attachment with reciprocity: properties and estimation. *Journal of Complex Networks*, 11(5):cnad031.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

de Melo Mendes, B. V. and Lopes, H. F. (2004). Data driven estimates for mixtures. *Computational Statistics & Data Analysis*, 47(3):583–598.

de Solla Price, D. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Dimitrova, D. S., Kaishev, V. K., and Tan, S. (2020). Computing the Kolmogorov-Smirnov distribution when the underlying CDF is purely discrete, mixed, or continuous. *Journal of Statistical Software*, 95(10):1–42.

Domenico, M. D., Solé-Ribalta, A., Gómez, S., and Arenas, A. (2014). Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences*, 111(23):8351–8356.

Eguíluz, V. M. (2005). Scale-free brain functional networks. *Physical Review Letters*, 94(1).

Erdös, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290.

Friedman, J. A. (2015). Using power laws to estimate conflict size. *Journal of Conflict Resolution*, 59(7):1216–1241.

Frigessi, A., Haug, O., and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3):219–235.

Gillespie, C. S. (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software*, 64(2):1–16.

Gillespie, C. S. (2017). Estimating the number of casualties in the American Indian war: A Bayesian analysis using the power law distribution. *The Annals of Applied Statistics*, 11(4).

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Hitz, A. S., Davis, R. A., and Samorodnitsky, G. (2024). Discrete extremes. *Journal of Data Science*, 22(4):524–536.

Holme, P. (2019). Rare and everywhere: Perspectives on scale-free networks. *Nature Communications*, 10(1).

Hu, Y. and Scarrott, C. (2018). evmix: An R package for extreme value mixture modeling, threshold estimation and boundary corrected kernel density estimation. *Journal of Statistical Software*, 84(5):1–27.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

Jung, H. and Phoa, F. K. H. (2021). A mixture model of truncated zeta distributions with applications to scientific collaboration networks. *Entropy*, 23(5):502.

Khanin, R. and Wit, E. (2006). How scale-free are biological networks. *Journal of Computational Biology*, 13(3):810–818. PMID: 16706727.

Kunegis, J. (2013). KONECT. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM.

Lee, C., Eastoe, E. F., and Farrell, A. (2024). Degree distributions in networks: Beyond the power law. *Statistica Neerlandica*, 78(4):702–718.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., and Åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908.

MacDonald, A., Scarrott, C., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157.

Mannion, S. and MacCarron, P. (2023). A robust method for fitting degree distributions of complex networks. *Journal of Complex Networks*, 11(4):cnad023.

Mehrabi, N., Morstatter, F., Peng, N., and Galstyan, A. (2019). Debiasing community detection: the importance of lowly connected nodes. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, pages 509–512. ACM.

Murphy, C., Tawn, J. A., and Varty, Z. (2025). Automated threshold selection and associated inference uncertainty for univariate extremes. *Technometrics*, 67(2):215–224.

Nakagawa, T. and Osaki, S. (1975). The discrete weibull distribution. *IEEE Transactions on Reliability*, R-24(5):300–301.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Otiniano, C. E. G., Gonçalves, C. R., and Dorea, C. C. Y. (2017). Mixture of extreme-value distributions: identifiability and estimation. *Communications in Statistics - Theory and Methods*, 46(13):6528–6542.

Pastor-Satorras, R. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119 – 131.

Prieto, F., Gómez-Déniz, E., and Sarabia, J. M. (2014). Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis & Prevention*, 71:38–49.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rohrbeck, C., Eastoe, E. F., Frigessi, A., and Tawn, J. A. (2018). Extreme value modelling of water-related insurance claims. *The Annals of Applied Statistics*, 12(1).

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10(1):33–60.

Shimura, T. (2012). Discretization of distributions in the maximum domain of attraction. *Extremes*, 15(3):299–317.

Slivkins, A. (2019). Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12(1–2):1–286.

So, M. K., Tiwari, A., Chu, A. M., Tsang, J. T., and Chan, J. N. (2020). Visualizing COVID-19 pandemic risk through network connectedness. *International Journal of Infectious Diseases*, 96:558–561.

Stein, M. L. (2021). A parametric model for distributions with flexible behavior in both tails. *Environmetrics*, 32(2):e2658.

Steinbock, C., Biham, O., and Katzav, E. (2019). Analytical results for the in-degree and out-degree distributions of directed random networks that grow by node duplication. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(8):083403.

Stumpf, M. P. H. and Porter, M. A. (2012). Critical truths about power laws. *Science*, 335:665 – 666.

Valero, J., Pérez-Casany, M., and Duarte-López, A. (2022). The Zipf-Polylog distribution: Modeling human interactions through social networks. *Physica A: Statistical Mechanics and its Applications*, 603:127680.

van der Hoorn, P., Voitalov, I., van der Hofstad, R., and Krioukov, D. (2020). Problems with classification, hypothesis testing, and estimator convergence in the analysis of degree distributions in networks. *arXiv preprint arXiv:2003.14012*.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv preprint arXiv:2102.00884*.

Voitalov, I., van der Hoorn, P., van der Hofstad, R., and Krioukov, D. (2019). Scale-free networks well done. *Phys. Rev. Res.*, 1:033034.

Wang, T. and Resnick, S. I. (2022). Asymptotic dependence of in- and out-degrees in a preferential attachment model with reciprocity. *Extremes*, 25(3):417–450.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Wolf, Y. I., Karev, G., and Koonin, E. V. (2002). Scale-free networks in biology: new insights into the fundamentals of evolution? *BioEssays*, 24(2):105–109.

Zhao, X., Scarrott, C., Oxley, L., and Reale, M. (2010). Extreme value modelling for forecasting market crisis impacts. *Applied Financial Economics*, 20(1-2):63–72.

Chapter 3

Conditional Extremes with Graphical Models

Conditional Extremes with Graphical Models

Abstract

Multivariate extreme value analysis quantifies the probability and magnitude of joint extreme events. Classical multivariate models, such as max-stable or multivariate generalised Pareto distributions, generally have a high computational cost of fitting, which limits their application. To overcome this, models based on the asymptotically dependent multivariate Pareto distribution have recently incorporated graphical models to induce sparsity and reduce the dimension of the parameter space. While this approach is computationally efficient, the assumption of asymptotic dependence is inappropriate for many applications. The conditional multivariate extreme value model (CMEVM) is a popular model for which the asymptotic dependence assumption is not required. Unfortunately, inference for this model is semi-parametric, and consequently, it has poor predictive performance in high dimensions. An extension of the CMEVM that allows both the incorporation and selection of sparse dependence structures, and fully parametric prediction is proposed. The approach fills a current gap in statistical methodology by extending graphical models to asymptotically independent multivariate extreme value models. To support inference in high dimensions, a stepwise inference procedure that is computationally efficient and loses no information or predictive power is proposed. Simulation studies show the model is highly flexible, and an application to discharges in the upper Danube River basin provides promising results.

3.1 Introduction

The development of statistical models to describe and predict multivariate extreme events is a crucial part of natural hazard risk assessment, especially for data arising from spatial, temporal, and spatio-temporal processes. For example, multivariate extreme value models have been adopted to predict risk from extreme snowfall (Blanchet and Davison, 2011), sea surface temperature (Simpson and Wadsworth, 2021), droughts (Oesting and Stein, 2018), river flows (Asadi et al., 2015; Keef et al., 2013), forest fires (Stephenson et al., 2015), precipitation (Westra and Sisson, 2011), wind-speed (Engelke et al., 2015), and ocean storms (Shooter et al., 2019), all of which exhibit complex behaviour, which can only be effectively captured by a flexible, multi-parameter model.

Central to multivariate extreme value modelling is the concept of extremal dependence. Let $V = \{1, \ldots, d\}$ and consider the d-dimensional absolutely continuous random vector $\boldsymbol{X} = \{X_j : j \in V\}$ with joint and marginal distributions $F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}[\boldsymbol{X} \leq \boldsymbol{x}]$ and $F_j(x_j) = \mathbb{P}[X_j \leq x_j]$, respectively. If the quantity $\chi_A := \lim_{u \to 1} \mathbb{P}[F_i(X_i) > u : i \in A]/(1-u)$ for

$A \subseteq V$ and $|A| \geq 2$ is strictly positive, then the components in A are likely to experience their extremes simultaneously and are said to show asymptotic dependence (AD) (Simpson et al., 2020). If $\chi_A = 0$ then the variables in A cannot be simultaneously extreme; specifically for $|A| = 2$, if $\chi_A = 0$ then the two variables show asymptotic independence (AI) (Ledford and Tawn, 1996). Full AD occurs when $\chi_A > 0$ for all subsets $A \subseteq V$, and full AI occurs when $\chi_A = 0$ for all two-dimensional subsets of V. While independence implies AI, the converse does not hold.

The strength of association between components with AI can be quantified using the coefficient of tail dependence η (Ledford and Tawn, 1996). The coefficient arises from a first-order approximation for the joint survivor function of (X_i, X_j) for $i, j \in A$, $i \neq j$. For large x the approximation is

$$\mathbb{P}[F_F^{-1}(F_{X_i}(X_i)) > x, F_F^{-1}(F_{X_j}(X_j)) > x] \sim \mathcal{L}(x)\mathbb{P}[F_F^{-1}(F_{X_i}(X_i)) > x]^{-\frac{1}{\eta}}, \tag{3.1.1}$$

where $F_F^{-1}(\cdot)$ is the inverse of the standard Fréchet distribution function and $\mathcal{L}(\cdot)$ is a slowly varying function. For pairs that exhibit AD, $\eta = 1$ and $\mathcal{L}(x) \nrightarrow 0$ as $x \to \infty$. Otherwise, X_i and X_j are either negatively ($0 < \eta < 0.5$) or positively ($0.5 < \eta < 1$) associated in their extremes, or exactly extremally independent ($\eta = 0.5$). Estimates of η are usually obtained over a range of finite thresholds, and the limit behaviour of $\eta(u)$, where u is the u-th quantile of the standard Fréchet distribution, is used to determine the likely extremal dependence class.

We illustrate these concepts using river discharges from the upper Danube River basin. Daily discharge data at $d = 31$ gauging stations for 1960-2009 is available from the Bavarian Environmental Agency (http://www.gkd.bayern.de). Figure 3.1 (left panel) shows the undirected tree implied by the flow connections of the river basin. We use the summer-only, temporally declustered dataset (see Asadi et al. (2015) for details), available from the `graphicalExtremes` package (Engelke et al., 2024b) in `R` (R Core Team, 2025). The data have previously been analysed using a max-stable Brown-Resnick process (Asadi et al., 2015) and a multivariate Pareto graphical model (Engelke and Hitz, 2020), both of which assume full AD. Figure 3.1 (right panel) shows scatter plots on standard Fréchet margins, and empirical estimates for the extremal dependence measure $\eta(u)$ (Ledford and Tawn, 1996) for stations 19 and 29 (top), and 19 and 16 (bottom). Sites 19 and 29 lie on different tributaries and are flow-unconnected, while 19 and 16 both lie on the Isar tributary and are flow-connected. Stations 19 and 16 appear to exhibit AD: they experience extreme events simultaneously and $\eta(u) \to 1$ as $u \to 1$. Conversely, the scatter plots for stations 19 and 29 suggest AI, which is supported by the fact that $\eta(u) \nrightarrow 1$ as $u \to 1$. Thus, while flow-connected stations often

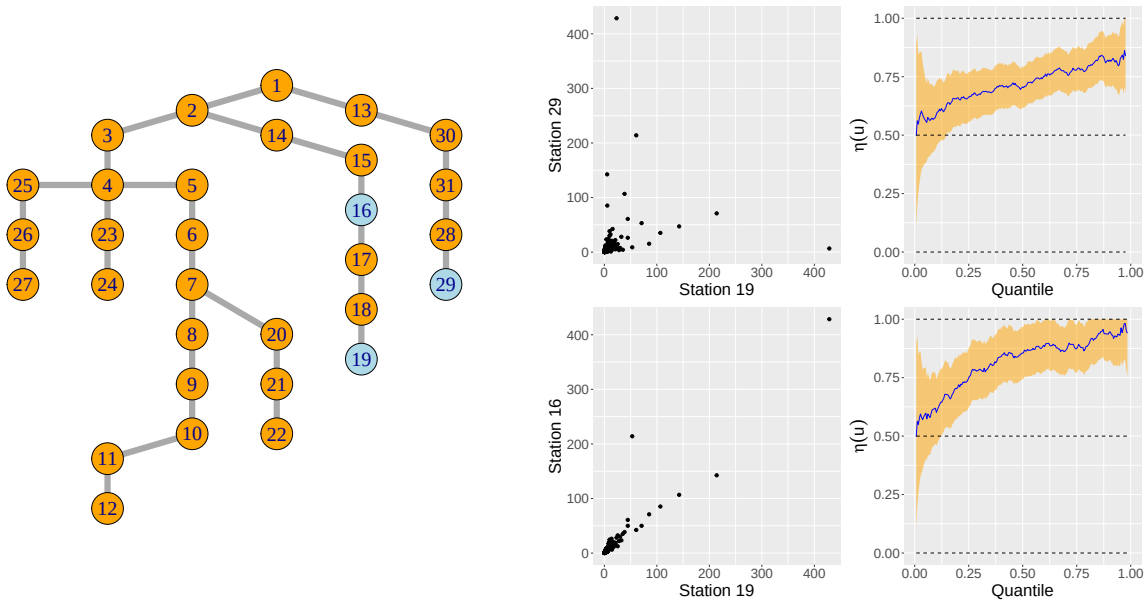

Figure 3.1: Undirected tree induced by the flow connections of the upper Danube River basin (left) with sites 16, 19 and 29 in blue. Scatter plots on standard Fréchet margins (centre) and empirical estimates of $\eta(u)$ (right) for $u \in (0, 1]$ for sites 19 and 29 (top) and 16 and 29 (right).

exhibit AD, some flow-unconnected stations do not and assuming AD for such stations would lead to overestimation of their joint tail behaviour.

Determining the class of extremal dependence is important, since not all multivariate extreme value models permit both AD and AI. Many popular models are based on max-stable distributions and processes, which are asymptotically dependent. Examples include the multivariate Pareto distribution (Rootzén and Tajvidi, 2006; Rootzén et al., 2018) and the max-stable (Smith, 1990) and generalised r-Pareto (Ferreira and de Haan, 2014) processes. The appeal of these comes from their connection to asymptotic limit theory, and is one of the reasons why the existing graphical modelling approach for extremes (Engelke and Hitz, 2020) is built on the Hüsler-Reiss distribution, which belongs to the class of multivariate Pareto distributions. While the original presentation in Engelke and Hitz (2020) allows only block graphs with cliques of size at most three, numerous subsequent extensions have been proposed (Engelke et al., 2024a). For example, Engelke et al. (2025) allow for *any* (sparse) graphical structure for the dependence structure, while Röttger et al. (2023) introduce coloured graphs which permit symmetries into the variogram matrix. However, the underlying distributional assumption makes these models unsuitable for data which exhibit AI.

Models for AI, such as the one proposed by Ledford and Tawn (1997), were initially limited to the bivariate case only. While Ramos and Ledford (2011) and Wadsworth and Tawn

(2013) developed extensions of this model that allowed for inference away from the diagonal, the semi-parametric conditional multivariate extreme value model (CMEVM) was the first model to provide a credible approach to data exhibiting both AD and AI (Heffernan and Tawn, 2004). The CMEVM is not based on a multivariate distribution or process; rather, conditional on one variable being large, normalising functions are defined to control the rate of growth of all other variables such that, after normalisation, the joint distribution of these "residuals" is non-degenerate. The model has gained popularity due to the relative ease with which its parameters can be estimated and interpreted, even in high dimensions. Applications include flood risk mapping (Neal et al., 2013; Towe et al., 2019), and the prediction of extreme sea states (Gouldby et al., 2014, 2017; Ross et al., 2020), sea surface temperatures (Simpson and Wadsworth, 2021), heatwaves (Winter and Tawn, 2016), and precipitation (Debusho and Diriba, 2021; Richards et al., 2022).

One issue with the CMEVM is that its predictive performance declines with increasing dimensionality (see next section and the Supplementary Material). The spatial CMEVM (Richards et al., 2022; Shooter et al., 2021; Wadsworth and Tawn, 2022) overcomes this by using a fully parametric spatial kernel for the residual distribution. By construction, this model is appropriate for measurements on a continuous spatial surface but not for measurements on topographies such as road or river networks, which are represented by a graph (Figure 3.1, left panel). To the best of our knowledge, the only work in this area is that of Papastathopoulos (2016) and Casey and Papastathopoulos (2023), who prove numerous important theoretical results for the CMEVM when processes are observed on decomposable graphs. While the authors develop an impressive theoretical framework, developing statistical methodology was beyond the scope of their contribution.

Our contribution fills this gap. We use the multivariate asymmetric generalised Gaussian (MVAGG) to model the CMEVM residuals, increasing accuracy for full parametric prediction. We also incorporate structure into the CMEVM residuals, thereby providing a framework for (sparse) graphical structures that accommodates both extremal dependence classes. The second of these contributions builds on ideas presented by Jennifer Wadsworth in the discussion of Engelke and Hitz (2020) and generalises the temporal Markov CMEVM of Winter and Tawn (2017). Finally, inference in high dimensions is achieved using stepwise optimisation that is computationally efficient without loss of information or predictive performance.

The remainder of this paper is structured as follows. Section 3.2 provides an overview of the CMEVM, introduces the MVAGG distribution, and describes the proposed structured CMEVM. Methods for model inference, graphical selection, and model-based predictions

are provided in Section 3.3. In Section 3.4, we illustrate the performance of our model, the graphical selection procedure, and the utility of the stepwise inference procedures. We then apply our model to discharges in the upper Danube River basin (Asadi et al., 2015) in Section 3.5 and compare it to the one proposed by Engelke and Hitz (2020). Finally, we outline directions for future research in Section 3.6.

3.2 Methodology

We review the CMEVM, introduce the MVAGG distribution, and describe a new variant of the CMEVM that incorporates sparsity into the residual distribution. Throughout, we use $V_{|i} := V \backslash \{i\}$ and $\boldsymbol{X}_{|i} := \{X_j : j \in V_{|i}\}$ to refer to the set V and vector $\boldsymbol{X}$ excluding their ith elements. Standard hat notation is used to denote parameter estimates.

3.2.1 Conditional Multivariate Extreme Value Model

Multivariate extreme value models are usually defined on either max-stable or heavy-tailed univariate margins. The use of specific margins is not restrictive since Sklar's Theorem (Sklar, 1959) allows one to transform the univariate margins of a random vector without altering the dependence structure. In what follows, $\boldsymbol{Y}$ and $\boldsymbol{Y}_{|i}$ denote the random vectors $\boldsymbol{X}$ and $\boldsymbol{X}_{|i}$ following transformation to Laplace margins. The CMEVM is based on a limiting representation of $\boldsymbol{Y}$ given a pre-selected conditioning component Y_i is extreme. The key innovation is the use of normalising functions which depend on the conditioning component. Specifically for each $i \in V$, suppose that there exist functions $\{a_{j|i} : \mathbb{R} \to \mathbb{R}, j \in V_{|i}\}$ and $\{b_{j|i} : \mathbb{R} \to \mathbb{R}_+, j \in V_{|i}\}$ such that as $u_{Y_i} \to \infty$,

$$\left(\left\{ \frac{Y_j - a_{j|i}(Y_i)}{b_{j|i}(Y_i)} \right\}_{j \in V_{|i}}, \; Y_i - u_{Y_i} \right) \Bigg| \; Y_i > u_{Y_i} \xrightarrow{d} \left(\{Z_{j|i} : j \in V_{|i}\}, \; E \right), \tag{3.2.1}$$

then the residual vector $\boldsymbol{Z}_{|i} = \{Z_{j|i} : j \in V_{|i}\}$ is has a non-degenerate distribution and is independent of the excesses of the conditioning component Y_i, which follow a standard exponential distribution E in the limit (Heffernan and Resnick, 2007).

To fit the CMEVM, limit (3.2.1) is assumed to hold exactly above a high but finite threshold u_{Y_i}. There are currently no methods for threshold selection. Therefore, sensitivity checks should be undertaken to ensure the threshold is low enough that the resulting estimates are reliable, but not so low that limit (3.2.1) does not hold. Once the threshold is selected, inference is undertaken separately on: (i) $Y_i - u_{Y_i} \mid Y_i > u_{Y_i}$; (ii) the normalising functions;

and (iii) the residuals $\boldsymbol{Z}_{|i}$. The first is trivial, since $Y_i - u_{Y_i} \mid Y_i > u_{Y_i}$ is standard exponential by limit (3.2.1). Parts (ii) and (iii) require greater consideration.

Heffernan and Tawn (2004) model the normalising functions as

$$a_{j|i}(y_i) = \alpha_{j|i} y_i, \quad b_{j|i}(y_i) = y_i^{\beta_{j|i}},$$

where for Laplace margins, $\alpha_{j|i} \in [-1, 1]$ and $\beta_{j|i} \in (-\infty, 1]$. These flexible functions capture AD ($\alpha_{j|i} = 1$ and $\beta_{j|i} = 0$), complete independence ($\alpha_{j|i} = 0$), and AI (all other parameter combinations). In contrast, there is no general class of distributions for modelling the residuals $\boldsymbol{Z}_{|i}$. Heffernan and Tawn (2004) use the working assumption that $\boldsymbol{Z}_{|i}$ follows a $(d-1)$-dimensional multivariate Gaussian (MVG) distribution, denoted $\boldsymbol{Z}_{|i} \sim \text{MVG}_{d-1}(\boldsymbol{\mu}_{|i}, \Sigma^*_{|i})$, with mean vector $\boldsymbol{\mu}_{|i} = \{\mu_{j|i} : j \in V_{|i}\} \in \mathbb{R}^{d-1}$ and covariance matrix $\Sigma^*_{|i}$. To further simplify, they take $\Sigma^*_{|i}$ to be diagonal so that the components of $\boldsymbol{Z}_{|i}$ are independent.

Inference is performed separately for each conditioning component Y_i. Under the assumption that the observations $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n$ are realisations of independent and identically distributed random vectors $\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^n$, the parameters of the normalising functions and the residual distribution associated with the ith conditioning component are obtained by maximising the likelihood

$$L_{|i}\left(\boldsymbol{\theta}_{|i}\right) = \prod_{k: y_i^k > u_{Y_i}} \phi_{d-1}\left[\left\{\frac{y_j^k - \alpha_{j|i} y_i^k}{\left(y_i^k\right)^{\beta_{j|i}}}\right\}_{j \in V_{|i}}; \boldsymbol{\mu}_{|i}, \Sigma^*_{|i}\right] \prod_{j \in V_{|i}} \left(y_i^k\right)^{-\beta_{j|i}}, \tag{3.2.2}$$

where $\phi_{d-1}(\cdot; \boldsymbol{\mu}_{|i}, \Sigma^*_{|i})$ is the density of the $(d-1)$-dimensional MVG distribution. The second term in the product is the Jacobian arising from the transformation of $\boldsymbol{Y}_{|i}$ to $\boldsymbol{Z}_{|i}$. While the MVG assumption is useful for parameter estimation, it is not necessarily an appropriate distributional assumption. Consequently, Heffernan and Tawn (2004) use the empirical distribution of the fitted residuals

$$\hat{\boldsymbol{z}}_{|i} = \{\hat{z}_{j|i} \coloneqq (y_{j|i} - \hat{\alpha}_{j|i} y_i) y_i^{-\hat{\beta}_{j|i}} : j \in V_{|i}\}, \tag{3.2.3}$$

to undertake model-based prediction.

3.2.2 Multivariate Asymmetric Generalised Gaussian Distribution

While the CMEVM is flexible and the computational cost of parameter (point) estimation is low, the predictive performance and interval estimation declines in high dimensions for the original model discussed in Section 3.2.1. This is due to the well-known curse of dimensionality associated with sampling from the empirical distribution (Nagler and Czado, 2016); see the Supplementary Material for an illustration in the specific case of the CMEVM. Overcoming this limitation requires a fully parametric model. While this approach has been taken before (Richards et al., 2022; Shooter et al., 2021; Wadsworth and Tawn, 2022), the model that we propose allows, for the first time, both asymmetry in the univariate marginal distributions of the residuals and a sparse dependence structure in their joint distribution.

The most commonly used fully parametric model for $\boldsymbol{Z}_{|i}$ is the MVG copula with generalised Gaussian margins (Wadsworth and Tawn, 2022). The generalised Gaussian distribution bridges the Gaussian and Laplace distributions, making it appropriate in cases where the pairwise dependence of the residuals is expected to vary from strong asymptotic dependence (for which the residual margins are expected to follow a Gaussian distribution) to complete independence (for which the residual margins are expected to follow a Laplace distribution). However, the generalised Gaussian distribution is symmetric, a property that we found is not always appropriate (see Supplementary Material). Instead, we model the marginal distribution of the residuals $\boldsymbol{Z}_{|i}$ using the asymmetric generalised Gaussian (AGG) distribution (Nacereddine and Goumeidane, 2019). This distribution has density

$$f_Z(z;\boldsymbol{\Theta}) = \frac{\delta}{(\kappa_1+\kappa_2)\Gamma(1/\delta)} \begin{cases} \exp\left\{-\left[(\nu-z)/\kappa_1\right]^\delta\right\} & z<\nu, \\ \exp\left\{-\left[(z-\nu)/\kappa_2\right]^\delta\right\} & z\geq\nu, \end{cases} \tag{3.2.4}$$

where $z \in \mathbb{R}$, $\Gamma(\cdot)$ denotes the standard gamma function, $\boldsymbol{\Theta} = (\nu, \kappa_1, \kappa_2, \delta)$, and $\nu \in \mathbb{R}$, $\kappa_1 > 0$, $\kappa_2 > 0$, $\delta > 0$ are the location, left-scale, right-scale, and shape parameters, respectively. We refer to this distribution as the AGG$(\nu, \kappa_1, \kappa_2, \delta)$. When $\kappa_1 = \kappa_2$, the AGG reduces to the generalised Gaussian (or delta-Laplace) distribution used by Wadsworth and Tawn (2022).

Any model for the marginal distribution of the residuals $\boldsymbol{Z}_{|i}$ must have two properties: (i) it can adapt its shape to account for the strength of extremal dependence, and (ii) it can account for asymmetry. As discussed, the generalised Gaussian distribution satisfies (i) but not (ii). Alternatively, distributions such as the skew normal, skew-t, and asymmetric Laplace distributions can capture (ii) but not (i). The AGG distribution is therefore an ideal

candidate as it can account for asymmetry in the margins while also containing the Gaussian (complete asymptotic dependence) and Laplace (complete independence) distributions as edge cases.

For inference, it is helpful to separate the marginal and dependence properties of the residual vector $\boldsymbol{Z}_{|i}$ by defining

$$W_{j|i} := \Phi^{-1}(F_{Z_{j|i}}(Z_{j|i})), \quad j \in V_{|i},$$

where Φ and $F_{Z_{j|i}}$ are the distribution functions of the standard Gaussian and $\text{AGG}(\nu, \kappa_1, \kappa_2, \delta)$ distributions, respectively. We can then assume that $\boldsymbol{W}_{|i} = \{W_{j|i} : j \in V_{|i}\} \sim \text{MVG}_{d-1}(\mathbf{0}, \Sigma_{|i})$ where $\Sigma_{|i}$ is a $(d-1)$-dimensional *correlation* matrix. We name this combination of marginal and joint distributions the multivariate asymmetric generalised Gaussian (MVAGG) distribution, denoted by $\boldsymbol{Z}_{|i} \sim \text{MVAGG}_{d-1}(\boldsymbol{\Theta}_{|i}, \boldsymbol{\Theta}^{\Gamma}_{|i})$, where $\boldsymbol{\Theta}_{|i} := \left\{(\nu_{j|i}, \kappa_{1_{j|i}}, \kappa_{2_{j|i}}, \delta_{j|i}) : j \in V_{|i}\right\}$ and $\boldsymbol{\Theta}^{\Gamma}_{|i}$ parameterises the (sparse) precision matrix $\Gamma_{|i} := (\Sigma_{|i})^{-1}$. The density for this distribution is

$$\begin{aligned} f_i(\boldsymbol{z}_{|i}; \boldsymbol{\Theta}_{|i}, \boldsymbol{\Theta}^{\Gamma}_{|i}) = \phi_{d-1} &\left[\left\{ \Phi^{-1}\left(F_{Z_{j|i}}\left(z_{j|i}; \boldsymbol{\Theta}_{j|i}\right)\right) \right\}_{j \in V_{|i}} ; \, \mathbf{0}, \Sigma_{|i} \right] \\ &\prod_{j \in V_{|i}} \frac{f_{Z_{j|i}}\left(z_{j|i}; \boldsymbol{\Theta}_{j|i}\right)}{\phi\left[\Phi^{-1}\left(F_{Z_{j|i}}\left(z_{j|i}; \boldsymbol{\Theta}_{j|i}\right)\right)\right]} \end{aligned} \tag{3.2.5}$$

where $f_{Z_{j|i}}$ is defined in equation (3.2.4), and ϕ is the standard univariate Gaussian density.

3.2.3 Structured Conditional Multivariate Extreme Value Model

While a parametric model for the residual component of the CMEVM permits prediction in high dimensions, it has the drawback of vastly increasing the number of model parameters. Specifically, the correlation matrix $\Sigma_{|i}$ that parameterises the MVAGG distribution increases the number of parameters from order d^2 (the original CMEVM) to order d^3. We now explain how conditional independence structures can overcome this challenge by introducing sparsity into the precision matrix $\Gamma_{|i}$ associated with $\Sigma_{|i}$.

We first introduce some necessary terminology. The conditional independence structure of a random vector $\boldsymbol{W} \sim \text{MVG}_d(\mathbf{0}, \Sigma)$ can be formulated by associating $\boldsymbol{W}$ with a simple

undirected graph $\mathcal{G} = (V, E)$. This graph consists of vertex $V = \{1, \ldots, d\}$ and edge $E \subseteq \{\{j, k\} \mid j, k \in V, j \neq k\}$ sets. The components W_j and W_k are conditionally independent given the remaining components if $\{j, k\} \notin E$. Since $\boldsymbol{W}$ additionally follows a Gaussian distribution, conditional independence of W_j and W_k implies that their partial correlation is zero, and hence that $\Gamma_{j,k} = \Gamma_{k,j} = 0$, where $\Gamma = (\Sigma)^{-1}$ is the precision matrix of $\boldsymbol{W}$ (Speed and Kiiveri, 1986). Thus, a sparse precision matrix (equivalently, a sparse graph $\mathcal{G}$) can greatly reduce the dimension of the parameter space.

To construct the structured CMEVM (SCMEVM), we begin by conditioning on a single component Y_i. Given that $Y_i > u_{Y_i}$, we define a conditional independence graph $\mathcal{G}_{|i} = \{E_{|i}, V_{|i}\}$ associated with the $(d-1)$-dimensional residual vector $\boldsymbol{W}_{|i}$. This graph might be defined by the topology on which the process is measured, or it might be estimated empirically. Assuming that the residuals $\boldsymbol{Z}_{|i}$ follow a MVAGG distribution, density (3.2.5) can be used to define the likelihood function such that the elements of the precision matrix $\Gamma_{|i}$ that correspond to edges not in the edge set $E_{|i}$ are set to zero. All other elements in the matrix are treated as free parameters to be estimated.

Conditioning in turn on each component results in d models, each with a graph $\mathcal{G}_{|i}$ that describes the dependence structure of the residual vector $\boldsymbol{W}_{|i}$. This presents a conundrum for model inference: how can these graphs be learnt? If we consider the limit result on which the CMEVM is constructed, all d conditional dependence graphs are inherited from the graph $\mathcal{G}_{\boldsymbol{X}}$ which represents the conditional dependence structure of the generating random vector $\boldsymbol{X}$. Using this observation, it seems reasonable to learn a unified graph from which the $\mathcal{G}_{|i}$ are subsequently inferred. In practice, $\mathcal{G}_{\boldsymbol{X}}$ is unknown and, following the extreme value paradigm of using only data in the tails to infer tail behaviour, we prefer not to estimate it directly. This leaves two options: ignore the asymptotic self-consistency between the graphs and infer each separately, or learn a unified graph from only the tail data. The former has the advantage of providing a more flexible modelling tool at the expense of increased computational cost. The latter ensures consistency across conditioning components at the expense of model flexibility. Scalability is a major motivation for our work, so we elect for the second approach.

Choosing a single graph presents two challenges: selection of the d-dimensional graph $\mathcal{G}$ and inference of the sub-graphs. The first will be addressed in Section 3.3.2. For the second, we augment $\boldsymbol{W}_{|i}$ to include $W_{i|i} = 0$, resulting in a d-dimensional vector. Exploiting textbook properties of the conditional MVG distribution, each $(d-1)$-dimensional precision matrix $\Gamma_{|i}$ can then be obtained by excluding the ith row and column from the matrix Γ associated with $\mathcal{G}$. Equivalently, since $\boldsymbol{W}$ follows a MVG distribution, the graph $\mathcal{G}_{|i}$ associated with $\boldsymbol{W}_{|i}$ is found by removing the ith node and its incident edges from $\mathcal{G}$. In further contrast to

learning the graphs individually, this approach allows *explicit* conditioning on $W_{i|i} = 0$.

3.3 Inference

In this section, we describe the inference scheme for the MVAGG SCMEVM. The methods described could be easily adapted to any other residual distribution based on a Gaussian copula. We also provide algorithms for graphical selection and model-based predictions.

3.3.1 Parameter estimation

Given n independent and identically distributed realisations $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n$ of the d-dimensional random vector $\boldsymbol{X}$, the first step is transformation to standard Laplace margins. By double application of the probability integral transform (PIT), for each $i \in V$ and each $k \in \{1, \ldots, n\}$,

$$y_i^k = \begin{cases} -\log\left(2\left[1 - \tilde{F}_i\left(x_i^k\right)\right]\right) & \tilde{F}_i\left(x_i^k\right) > 0.5, \\ \log\left(2\tilde{F}_i\left(x_i^k\right)\right) & \tilde{F}_i\left(x_i^k\right) \leq 0.5, \end{cases} \tag{3.3.1}$$

where $\tilde{F}_i$ is an estimate of the marginal distribution F_i for X_i. We use a semi-parametric estimate for F_i consisting of the empirical distribution for $x_i \leq v_{X_i}$ and a generalised Pareto distribution for $x_i > v_{X_i}$ (Heffernan and Tawn, 2004). The threshold v_{X_i} is selected using the automated method of Murphy et al. (2025).

Inference for the MVAGG SCMEVM is performed for each component Y_i separately by maximising the likelihood

$$L_{|i}\left(\boldsymbol{\theta}_{|i}\right) = \prod_{k:y_i^k > u_{Y_i}} f_i\left(\left\{\frac{y_j^k - \alpha_{j|i} y_i^k}{\left(y_i^k\right)^{\beta_{j|i}}}\right\}_{j \in V_{|i}}; \boldsymbol{\Theta}_{|i}, \boldsymbol{\Theta}_{|i}^{\Gamma}\right) \prod_{j \in V_{|i}} \left(y_i^k\right)^{-\beta_{j|i}}, \tag{3.3.2}$$

where f_i is given by equation (3.2.5), u_{Y_i} is the dependence threshold and $\boldsymbol{\theta}_{|i} := (\boldsymbol{\Theta}_{|i}^d, \boldsymbol{\Theta}_{|i}, \boldsymbol{\Theta}_{|i}^{\Gamma})$ combines the CMEVM dependence $\boldsymbol{\Theta}_{|i}^d := \left\{(\alpha_{j|i}, \beta_{j|i}) : j \in V_{|i}\right\}$, MVAGG marginal $\boldsymbol{\Theta}_{|i}$ and MVAGG correlation $\boldsymbol{\Theta}_{|i}^{\Gamma}$ parameters. Of the parameter vectors, only $\boldsymbol{\Theta}_{|i}^{\Gamma}$ is determined by the graph associated with $\boldsymbol{W}_{|i}$. The edge cases are the independent model, a graph with no edges, and the saturated model, a full graph. In the former, only the diagonal of the precision matrix is estimated, with all other elements set to zero. In the latter, all elements of the precision matrix are estimated. Any other combination will be referred to as a graphical model.

Algorithm 3.1 One-step parameter estimation for the MVAGG SCMEVM

1: Initialise $\boldsymbol{\Theta}_{|i}^{d}$, $\boldsymbol{\Theta}_{|i}$, $\mathcal{G}_{|i} = (V_{|i}, E_{|i})$, and ε;
2: The current values of $\boldsymbol{\Theta}_{|i}^{d}$ and $\boldsymbol{\Theta}_{|i}$ are $\boldsymbol{\Theta}_{|i}^{d*}$ and $\boldsymbol{\Theta}_{|i}^{*}$, respectively;
3: Obtain $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma} = \underset{\boldsymbol{\Theta}_{|i}^{\Gamma}}{\operatorname{argmax}} L_{|i}\left(\boldsymbol{\Theta}_{|i}^{d*}, \boldsymbol{\Theta}_{|i}^{*}, \boldsymbol{\Theta}_{|i}^{\Gamma}\right)$, where $L_{|i}$ is likelihood (3.3.2) using Algorithm 3.2;
4: Obtain $\left(\hat{\boldsymbol{\Theta}}_{|i}^{d}, \hat{\boldsymbol{\Theta}}_{|i}\right) = \underset{\boldsymbol{\Theta}_{|i}^{d}, \boldsymbol{\Theta}_{|i}}{\operatorname{argmax}} L_{|i}\left(\boldsymbol{\Theta}_{|i}^{d}, \boldsymbol{\Theta}_{|i}, \hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}\right)$;
5: **if** $||\ (\hat{\boldsymbol{\Theta}}_{|i}^{d}, \hat{\boldsymbol{\Theta}}_{|i}), (\boldsymbol{\Theta}_{|i}^{d*}, \boldsymbol{\Theta}_{|i}^{*})\ ||> \varepsilon$ **then**
6: Set $\boldsymbol{\Theta}_{|i}^{d*} = \hat{\boldsymbol{\Theta}}_{|i}^{d}$ and $\boldsymbol{\Theta}_{|i}^{*} = \hat{\boldsymbol{\Theta}}_{|i}$;
7: Repeat steps 3 - 4;
8: **else**
9: **return** $\hat{\boldsymbol{\Theta}}_{|i}^{d}$, $\hat{\boldsymbol{\Theta}}_{|i}$ and $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}$
10: **end if**

The naive approach is to jointly estimate the full parameter vector $\boldsymbol{\theta}_{|i}$. A "one-step" numerical maximisation procedure for this is given in Algorithm 3.1. The algorithm iterates between maximising the profile likelihood for $\boldsymbol{\Theta}_{|i}^{\Gamma}$ and maximising the profile likelihood for $(\boldsymbol{\Theta}_{|i}^{d}, \boldsymbol{\Theta}_{|i})$. In step 4 of Algorithm 3.1, $(\boldsymbol{\Theta}_{|i}^{d}, \boldsymbol{\Theta}_{|i})$ are maximised using the BFGS quasi-Newton method, so the norm used in step 5 is the sum of squares. Note, the parameters for $\boldsymbol{\Theta}_{|i}^{\Gamma}$ need not be included in this condition since the profile likelihood for $\boldsymbol{\Theta}_{|i}^{\Gamma}$ has previously been maximised at the current values $(\boldsymbol{\Theta}_{|i}^{d*}, \boldsymbol{\Theta}_{|i}^{*})$ using Algorithm 3.2.

Numerical optimisation of the profile likelihood for $\boldsymbol{\Theta}_{|i}^{\Gamma}$ is detailed in Algorithm 3.2 and uses the input graph $\mathcal{G}_{|i}$ to determine the precision matrix $\Gamma_{|i}$. Specifically, numerical optimisation is only required for non-trivial graphical structures since a closed-form expression exists for $\hat{\Sigma}_{|i}$, and hence also for $\hat{\Gamma}_{|i}$, for both the independent and saturated models, i.e. the identity and inverse correlation matrix, respectively. For the graphical model, the graph $\mathcal{G}_{|i}$ is chosen *a priori*; see Section 3.3.2 for details. As discussed, if $\{j, k\} \notin E_{|i}$ then $(\hat{\Gamma}_{|i})_{j,k}$ must be 0, a condition which can be enforced by using the graphical lasso (Friedman et al., 2007, Remark 2.1) to estimate $\Gamma_{|i}$. This is the only complex step in the algorithm, however, provided that the underlying data, $\boldsymbol{W}_{|i}$, is Gaussian and the dependence structure is correctly specified by the graph $\mathcal{G}_{|i}$, there should be few issues.

Although Algorithm 3.1 should converge, as it is iterating between maximising profile likelihoods and can be seen as adjacent to other optimisation techniques such as conjugate gradient methods (Shewchuk et al., 1994) and Gibbs sampling (Yildirim, 2012), the procedure has limitations. For example, the estimates of $\boldsymbol{\Theta}_{|i}^{d}$ and $\boldsymbol{\Theta}_{|i}$ are not independent: the estimate of $\alpha_{j|i}$ ($\beta_{j|i}$) influences the mode (variance) of the residual distribution. Consequently, while joint

Algorithm 3.2 Estimating $\boldsymbol{\Theta}_{|i}^{\Gamma}$

Initialise $\hat{\boldsymbol{\Theta}}_{|i}^{d}$, $\hat{\boldsymbol{\Theta}}_{|i}$, and $\mathcal{G}_{|i} = (V_{|i}, E_{|i})$;
Obtain $\hat{\boldsymbol{z}}_{|i}$ using equation (3.2.3) and $\hat{\boldsymbol{\Theta}}_{|i}^{d}$;
Obtain $\hat{\boldsymbol{w}}_{|i}$ such that $\hat{w}_{j|i} = \Phi^{-1}(F_{Z_{j|i}}(\hat{z}_{j|i}; \hat{\boldsymbol{\Theta}}_{j|i}))$ for $j \in V_{|i}$;
if $|E_{|i}| = 0$ **then** *Independence*
    $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma} = I_{d-1}$ (the $(d-1)$-dimensional identity matrix);
else if $|E_{|i}| = d(d-1)/2$ **then** *Saturated*
    $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma} = (\text{corr}(\hat{\boldsymbol{W}}_{|i}))^{-1}$;
else *Graphical*
    $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}$ is estimated using a graphical lasso (Friedman et al., 2019) on $\hat{\boldsymbol{W}}_{|i}$;
return $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}$

estimation may result in a model with good predictive abilities (see Supplementary Material), the first-order extremal dependence structure is not entirely captured by the dependence parameters $\boldsymbol{\Theta}_{|i}^{d}$. Further, since AD is an edge case in the parameter space, this procedure is more likely to suggest that pairs of variables are AI when they are AD. Therefore, convergence of the algorithm should be investigated more thoroughly. Another limitation is that finding suitable initial values for the numerical optimisation becomes increasingly difficult for large d. Even for a sparse precision matrix $\Gamma_{|i}$, the parameter space grows at least linearly in d.

Thus, the one-step approach is only applicable in low dimensions, and even then, it is not recommended due to potential offsetting between the CMEVM dependence and MVAGG parameters. To address these issues, two- and three-step estimation procedures are described in Algorithm 3.3 and Algorithm 3.4, respectively.

The two-step approach in Algorithm 3.3 first estimates the SCMEVM dependence parameters $\boldsymbol{\Theta}_{|i}^{d}$ using the original CMEVM approach, i.e. we assume the residuals are independently Gaussian. Treating these parameters as fixed, similar to a plug-in estimator, we obtain the fitted residuals and then fit the MVAGG in equation (3.2.5) by again iterating between maximising the profile likelihood for $\boldsymbol{\Theta}_{|i}^{\Gamma}$ and maximising the profile likelihood for $\boldsymbol{\Theta}_{|i}$. This method is attractive because it ensures that the first-order extremal dependence structure is *entirely* captured by the dependence parameters $\boldsymbol{\Theta}_{|i}^{d}$. Convergence of the algorithm should be reviewed due to the plug-in nature of the dependence parameters $\boldsymbol{\Theta}_{|i}^{d}$. However, the separation of $\boldsymbol{\Theta}_{|i}^{d}$ and $\boldsymbol{\Theta}_{|i}$ means the former will converge at the same rate as the original CMEVM (Heffernan and Resnick, 2007) and the latter converges at the rate of a Gaussian distribution since it is a Gaussian copula.

Algorithm 3.3 Two-step parameter estimation for the MVAGG SCMEVM

1: Initialise $\boldsymbol{\Theta}_{|i}^{d}$, $\boldsymbol{\Theta}_{|i}$, $\mathcal{G}_{|i} = (V_{|i}, E_{|i})$, and ε;
2: Assuming independent Gaussian residuals, obtain $\hat{\boldsymbol{\Theta}}_{|i}^{d}$ by maximising likelihood (3.2.2);
3: Using equation (3.2.3) and $\hat{\boldsymbol{\Theta}}_{|i}^{d}$, obtain $\hat{\boldsymbol{z}}_{|i}$ and treat them as fixed;
4: The current value of $\boldsymbol{\Theta}_{|i}$ is $\boldsymbol{\Theta}_{|i}^{*}$;
5: Obtain $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma} = \underset{\boldsymbol{\Theta}_{|i}^{\Gamma}}{\operatorname{argmax}} L_{|i}\left(\hat{\boldsymbol{\Theta}}_{|i}^{d}, \boldsymbol{\Theta}_{|i}^{*}, \boldsymbol{\Theta}_{|i}^{\Gamma}\right)$, where $L_{|i}$ is likelihood (3.3.2) using Algorithm 3.2;
6: Obtain $\hat{\boldsymbol{\Theta}}_{|i} = \underset{\boldsymbol{\Theta}_{|i}}{\operatorname{argmax}} L_{|i}\left(\hat{\boldsymbol{\Theta}}_{|i}^{d}, \boldsymbol{\Theta}_{|i}, \hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}\right)$;
7: **if** $|| \hat{\boldsymbol{\Theta}}_{|i}, \boldsymbol{\Theta}_{|i}^{*} ||> \varepsilon$ **then**
8: set $\boldsymbol{\Theta}_{|i}^{*} = \hat{\boldsymbol{\Theta}}_{|i}$;
9: repeat 5 - 6;
10: **else**
11: **return** $\hat{\boldsymbol{\Theta}}_{|i}$ and $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}$.
12: **end if**
13: **return** $\hat{\boldsymbol{\Theta}}_{|i}^{d}$, $\hat{\boldsymbol{\Theta}}_{|i}$ and $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}$

The two-step approach still requires a computationally expensive numerical optimisation procedure for the $(d-1)$-dimensional MVAGG distribution and so is only applicable in low dimensions. To circumvent this, we propose a three-step approach in Algorithm 3.4. Rather than maximising the full $(d-1)$-dimensional MVAGG distribution, we separate estimation of the margins and dependence structure. This results in a much smaller parameter space to maximise over, leading to computational gains over the two-step approach and making it scalable to high dimensions. Furthermore, separating inference in this manner should not impact the parameter estimates or convergence since the MVAGG distribution is a Gaussian copula. Thus, we also don't lose information or convergence properties. Hence, the three-step approach is our preferred method.

We make two final observations. Firstly, in contrast to the stationary spatial CMEVM (Richards et al., 2022; Wadsworth and Tawn, 2022), in the SCMEVM, the parameter values differ with the conditioning variable. Hence, there is no information to be gained by jointly fitting the d conditional models. Secondly, by construction, Algorithm 3.4 avoids maximising the joint likelihood. Consequently, we cannot obtain uncertainty estimates using the standard asymptotic properties of the likelihood and the maximum likelihood estimators. To obtain these, we recommend using a non-parametric bootstrapping algorithm. This requires sampling with replacement from the original data to create artificial datasets, fitting the model to each dataset, and hence obtaining a bootstrap approximation to the sampling distributions of parameter estimates and model predictions.

Algorithm 3.4 Three-step parameter estimation for the MVAGG SCMEVM

1: Initialise $\boldsymbol{\Theta}_{|i}^{d}$, $\boldsymbol{\Theta}_{|i}$, and $\mathcal{G}_{|i} = (V_{|i}, E_{|i})$;
2: Assuming independent Gaussian residuals, obtain $\hat{\boldsymbol{\Theta}}_{|i}^{d}$ by maximising likelihood (3.2.2);
3: Using equation (3.2.3) and $\hat{\boldsymbol{\Theta}}_{|i}^{d}$, obtain $\hat{\boldsymbol{z}}_{|i}$ and treat them as fixed;
4: Assuming the components of $\hat{\boldsymbol{Z}}_{|i}$ are independent, obtain $\hat{\boldsymbol{\Theta}}_{|i} = \underset{\boldsymbol{\Theta}_{|i}}{\operatorname{argmax}}\, f_i(\hat{\boldsymbol{z}}_{|i}; \boldsymbol{\Theta}_{|i}, I_{d-1})$ where f_i is given by equation (3.2.5), and I_{d-1} is a $(d-1)$-dimensional identity matrix;
5: Obtain $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma} = \underset{\boldsymbol{\Theta}_{|i}^{\Gamma}}{\operatorname{argmax}}\, L_{|i}\left(\hat{\boldsymbol{\Theta}}_{|i}^{d}, \hat{\boldsymbol{\Theta}}_{|i}, \boldsymbol{\Theta}_{|i}^{\Gamma}\right)$, where $L_{|i}$ is likelihood (3.3.2) using Algorithm 3.2;
6: **return** $\hat{\boldsymbol{\Theta}}_{|i}^{d}$, $\hat{\boldsymbol{\Theta}}_{|i}$ and $\hat{\boldsymbol{\Theta}}_{|i}^{\Gamma}$

3.3.2 Graph selection

We now discuss selection of the graphs $\mathcal{G}_{|1}, \ldots, \mathcal{G}_{|d}$. Recall that we assume that these graphs are all derived from a unifying graph $\mathcal{G}$. In most cases, $\mathcal{G}$ will be unknown and must be learnt. Several such learning algorithms have been proposed for multivariate generalised Pareto distributions. Engelke and Hitz (2020) iteratively add edges to $\mathcal{G}$ to minimise the AIC, but this is costly in higher dimensions. Wan and Zhou (2025) present the extremal graphical lasso, an extension of the graphical lasso (Friedman et al., 2007; Yuan and Lin, 2007), while Engelke et al. (2025) propose "EGlearn", which combines a majority rule with either the graphical lasso or neighbourhood selection (Meinshausen and Bühlmann, 2006).

For consistency with Algorithm 3.2, our approach, detailed in Algorithm 3.5, also uses the graphical lasso and a majority rule. The algorithm has two tuning parameters: the thresholds $u_{Y_1}, \ldots, u_{Y_d}$ and the majority rule proportion p. The graphical lasso penalty parameter λ is not a tuning parameter as it is selected objectively by comparing the composite AIC scores. While other metrics are available, the AIC gives the graph with the best predictive properties. Returning to the tuning parameters, the thresholds should not be so low that limit (3.2.1) is a poor approximation; at the same time, if they are too high, there will be insufficient data to accurately identify the conditional dependence structure. The choice of the majority rule proportion p results in a similar trade-off: too high a value of p risks inferring a very sparse structure that predicts poorly; too low a value of p infers a dense graph that is computationally expensive to work with. In our applications, we set $p = 0.5$. We chose this because the inferred structure is not sensitive to the choice of p (see Supplementary Material) and is consistent with other majority rules used in extremal graphical selection (Engelke and Hitz, 2020; Engelke et al., 2025).

Algorithm 3.5 requires us to first set the majority rule proportion p and the SCMEVM

Algorithm 3.5 Graphical selection using the MVAGG SCMEVM

1: Initialise $\boldsymbol{\lambda}$, p and $u_{Y_1}, \ldots, u_{Y_d}$.
2: **for** $j = 1, \ldots, |\boldsymbol{\lambda}|$ **do**
3: **for** $i = 1, \ldots, d$ **do**
4: Assuming independent Gaussian residuals, obtain $\hat{\boldsymbol{\Theta}}^d_{|i}$ by maximising likelihood (3.2.2) with threshold u_{Y_i};
5: Using equation (3.2.3) and $\hat{\boldsymbol{\Theta}}^d_{|i}$, obtain $\hat{\boldsymbol{z}}_{|i}$ and treat them as fixed;
6: Assuming the components of $\hat{\boldsymbol{Z}}_{|i}$ are independent, obtain $\hat{\boldsymbol{\Theta}}_{|i} = \underset{\boldsymbol{\Theta}_{|i}}{\operatorname{argmax}}\, f_i(\hat{\boldsymbol{z}}_{|i}; \boldsymbol{\Theta}_{|i}, I_{d-1})$ where f_i is given by equation (3.2.5) and I_{d-1} is a $(d-1)$-dimensional identity matrix;
7: Set $\boldsymbol{\Theta}_{|i} = \hat{\boldsymbol{\Theta}}_{|i}$ and treating as fixed marginally transform $\hat{\boldsymbol{z}}_{|i}$ onto standard Gaussian margins $\hat{\boldsymbol{w}}_{|i}$;
8: Apply a graphical lasso with penalisation parameter λ_j to $\hat{\boldsymbol{W}}_{|i}$ to infer $\mathcal{G}_{|i}$;
9: **end for**
10: Obtain a weighted graph $\mathcal{G}^*$ by combining the subgraphs $\mathcal{G}_{|i}$;
11: Create $\mathcal{G}'$ by pruning the edges of $\mathcal{G}^*$ that do not occur at least $(p \times 100)\%$ of the time;
12: **for** $i = 1, \ldots, d$ **do**
13: Maximise likelihood (3.3.2) using Algorithm 3.4 and $\mathcal{G}'_{|i}$ obtained by removing the ith node and its incident edges from $\mathcal{G}'$;
14: **end for**
15: Calculate and store the composite AIC
16: **end for**
17: **return** $\mathcal{G}'$ that minimises the composite AIC from step 15.

dependence thresholds u_{Y_i} for each $i \in V$. We also need to specify the penalisation parameters $\boldsymbol{\lambda} \in (0, 1)$ that will be used in the graphical lasso. For each penalisation parameter λ, we obtain the fitted residuals on Gaussian margins for each conditioning site $i \in V$. This is achieved by first obtaining the fitted residuals $\boldsymbol{Z}_{|i}$ using the original CMEVM, and then transforming them onto Gaussian margins $\boldsymbol{W}_{|i}$ using marginally fitted AGG models under the assumption the components of $\boldsymbol{Z}_{|i}$ are independent. This is consistent with the three-step inference procedure proposed in Algorithm 3.4. Now, we can infer $\mathcal{G}_{|i}$ for each $i \in V$ using the graphical lasso with the specified penalisation parameter. As discussed, these graphs may not necessarily be consistent. Thus, we combine them into a single weighted graph $\mathcal{G}^*$ and prune the edges that do not occur at least $(p \times 100)\%$ of the time to create $\mathcal{G}'$. The idea is that $\mathcal{G}'$ is the best average graph over the d conditioning sites. We then refit the d conditioning models using the three-step inference procedure, assuming the dependence structure is consistent with $\mathcal{G}'$. These d model fits are used to calculate d AIC scores, which we combine to give a single composite AIC score. We repeat this for each penalisation parameter λ, and return the value of λ and its associated graph $\mathcal{G}'$ that minimises the composite AIC score. Note

that the composite AIC scores cannot be treated as a true composite score, since they do not account for the overlap of information in the d models. However, the same data is used over each penalisation parameter λ, since the dependence thresholds u_{Y_i} for $i \in V$ are fixed. Thus, comparing the scores in this manner again gives the best average graph over the d conditioning sites.

3.3.3 Prediction

By construction, the CMEVM and, by extension, the SCMEVM do not permit closed forms for either tail probabilities or quantiles. Heffernan and Tawn (2004) use a simulation-based prediction algorithm based on the empirical distribution of the fitted residuals $\hat{\boldsymbol{Z}}_{|i}$. A key motivation for introducing the SCMEVM in Section 3.2 was the observation that this prediction procedure fails in high dimensions. We now explain how the SCMEVM can be used to obtain fully parametric predictions using a method very similar to Wadsworth and Tawn (2022, Section 5.2.2) and Richards et al. (2022, Section 3.3).

For $u > \max(u_{Y_i} : i \in V)$, the SCMEVM describes the distribution of $\boldsymbol{X}$ given that the largest component of $\boldsymbol{Y}$ exceeds u, that is

$$\left\{ \tilde{F}_{X_i}^{-1}(F_L(Y_i)) : i \in V \right\} \,\middle|\, \left(\max_{i \in V} Y_i > u \right), \tag{3.3.3}$$

where u_{Y_i} is the SCMEVM dependence threshold for conditioning component Y_i, F_L is the distribution function of the standard Laplace distribution, and $\tilde{F}_{X_i}$ is the estimated marginal distribution of X_i. To create realisations of $\boldsymbol{X}$, we draw samples from equation (3.3.3) using Algorithm 3.6 with probability

$$\mathbb{P}\left(\max_{i \in V} Y_i > u \right) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\left\{ \max_{i \in V} y_i^k > u \right\}.$$

Otherwise, we draw realisations from the empirical distribution of $\boldsymbol{X} \,\middle|\, \left(\max_{i \in V} Y_i < u \right)$.

Algorithm 3.6 can be explained as follows. First, we set a threshold $u > \max(u_{Y_i} : i \in V)$ above which to simulate our extreme events. Next, we randomly select a conditioning site $i \in V$ and simulate the excesses for this site from a standard Exponential distribution due to limit (3.2.1). For the dependent site, we simulate these from the MVAGG in equation (3.2.5), which we achieve by simulating from a $(d-1)$-dimensional multivariate Gaussian distribution with correlation matrix $\Sigma_{|i}$ and marginally transforming onto AGG margins using $\boldsymbol{\Theta}_{|i}$. These are then transformed onto standard Laplace margins using limit (3.2.1) and the previously

Algorithm 3.6 Simulating data with at least one extreme event

Initialise u;
for $l = 1, \dots N$ such that $N > n$ **do**
  Draw a conditioning random variable i from $i \in V$ with uniform probability;
  Simulate $E^l \sim \text{Exp}(1)$ and set $y_i^l = u + E^l$;
  Simulate $\boldsymbol{z}_{|i}^l$ from the distribution described in Section 3.2.2;
  Calculate $y_{j|i}^l = \hat{\alpha}_{j|i} y_i^l + \left(y_i^k\right)^{\hat{\beta}_{j|i}} z_{j|i}^l$ for $j \in V_{|i}$;
  Calculate an importance weight $w^l = \left(\frac{1}{d}\sum_{m=1}^{d} \mathbb{1}\{y_m^l > u\}\right)^{-1}$.
end for
Sub-sample n realisations from $\{\boldsymbol{y}^1, \dots, \boldsymbol{y}^N\}$ with probabilities proportional to their importance weights;
Transform the sub-sample $\{\boldsymbol{y}^1, \dots, \boldsymbol{y}^n\}$ to $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^n\}$ via a double application of the PIT;
return $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^n\}$

simulated excesses. These are combined with the excesses to obtain a simulated vector of $\boldsymbol{Y} \mid Y_i > u$. We repeat this N times. We subsample n of these N vectors according to some importance weights. The weights aim to upweight/downweight samples with few/many extreme events so that samples with many extreme events are not overrepresented in the simulated data. In a spatial setting, this can be interpreted as downweighting samples in the centre of the domain and upweighting those towards the edge of the domain. These n samples are then transformed back onto the scale of the original data using the inverse of equation (3.3.1).

3.4 Simulation Study

In this section, we use simulation studies to assess the performance of the SCMEVM. We compare the three stepwise inference procedures and assess the graphical selection process. Finally, we compare the SCMEVM to existing methods.

3.4.1 Stepwise inference procedures

We consider the 5-dimensional SCMEVM with dependence structure given by the graph $\mathcal{G}$ with edge set $E = \{\{1,2\},\{1,3\},\{2,3\},\{3,4\},\{3,5\},\{4,5\}\}$. Using data simulated from this *true* model, we compare the performance of the one-(Algorithm 3.1), two-(Algorithm 3.3), and three-step (Algorithm 3.4) estimation procedures for each of three candidate dependence structures: independent, graphical, and saturated.

For each $i \in V$, the conditioning variable $Y_i \mid Y_i > u_{Y_i}$ is simulated from a standard Laplace

distribution with u_{Y_i} the 0.80-quantile of this distribution. To simulate the residual vector $\boldsymbol{Z}_{|i}$ from the MVAGG, the PIT is used to transform the margins of $\boldsymbol{W}_{|i}$, which are drawn from a MVG with standard margins and correlation matrix $\Sigma_{|i}$. Finally, the vector $\boldsymbol{Y}_{|i}$ is obtained by applying the inverse normalisation of limit (3.2.1). We generate 200 samples of $\boldsymbol{Y} \mid Y_i > u_{Y_i}$ and consider $n \in \{250, 500\}$. The true parameters are independently sampled from a uniform distribution on $(0.1, 0.5)$ for α_j, $(0.1, 0.3)$ for β_j, $(-5, 5)$ for ν_j, $(0.5, 2)$ for κ_{1_j}, $(1.5, 3)$ for κ_{2_j}, and $(0.8, 2.5)$ for δ_j, for each $j \in V$. Correlations in $\Sigma_{|i}$ correspond to weak positive associations; see Supplementary Material for the cases of weak negative and strong positive associations.

Note, since we are simulating data directly from the limiting distribution, the model should converge to the true parameter estimates almost instantly. Thus, we do not require a large sample size to assess the bias. However, the small sample size may impact the variability of the estimates. Also note that data are sampled from each component, yielding five unique datasets for analysis. We then condition on the corresponding component being large in the subsequent model fitting. To obtain a single dataset in which at least one component is large, we could use importance sampling (similar to Algorithm 3.6). However, this complicates the analysis of sensitivity to sample size, so we do not investigate this here. We choose the true model in this manner to (i) allow for comparisons between the different stepwise fitting procedures (a larger dimension would have made fitting the one-step model numerically difficult), (ii) the graphical structure is simple but offers conditional independence properties to highlight when certain models would be inappropriate, (iii) the range of correlation options allows us to assess if the model is underperforming in certain scenarios, and (iv) the parameter values should exhibit a range of behaviour in the first-order extremal dependence structure as well as in the residual distribution. The Supplementary Material provides more detail on each of these.

By construction, the estimates of $\boldsymbol{\Theta}_{|i}^d$ are the same under the two- and three-step procedures, regardless of the residual dependence structure. Consequently, we only compare $\hat{\boldsymbol{\Theta}}_{|i}^d$ for the one-step (all three dependence structures) and two-step (independence only) procedures. Figure 3.2 shows the bias in $\hat{\boldsymbol{\alpha}}_{|i}$, with results for the other parameters found to be similar (see Supplementary Material). Reassuringly, even when using the stepwise procedure, the true parameter values are recovered. The biases from the one-step method do have a slightly narrower range, while the two-step method has fewer instances of unusually large bias. As expected, the variability in bias decreases as the sample size increases for both procedures.

Table 3.1 presents the biases of the fitted maximum log-likelihood values. Although the

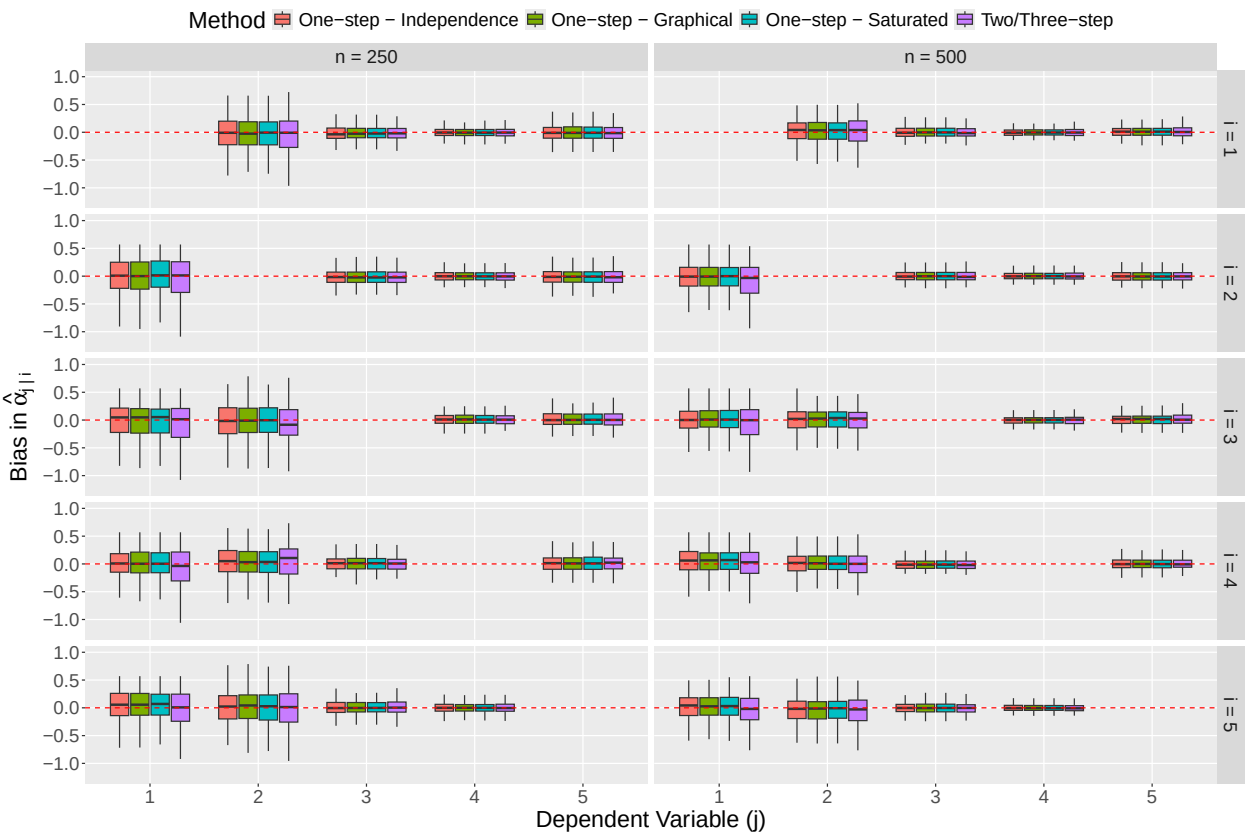

Figure 3.2: Boxplots detailing the bias of $\hat{\alpha}_{j|i}$ for distinct $i, j \in V$. Each row corresponds to the conditioning variable i, and each column corresponds to the sample size. The fill of the boxplots denotes the different models. The red dashed line indicates the $y = 0$ line.

SCMEVMs with graphical or saturated covariance exhibit slightly higher bias than the independent residuals model, the bias is consistent across all three stepwise procedures. Furthermore, we observe that the bias increases with sample size for the independent residual models, while for the graphical and saturated residual models, it remains similar for both $n = 250$ and $n = 500$. This suggests that the structured models are more robust to sample size changes.

Having assessed the accuracy of the stepwise procedures, we now evaluate computational efficiency. We consider $n \in \{250, 500, 1000, 2000, 4000\}$ excesses above the dependence threshold and dimensions $d \in \{5, 10, 15\}$. For each combination of sample size and dimension, a single sample is drawn from the SCMEVM. For comparison across dimensions, the number of edges in each graph is set to 60% of the maximum possible number of edges. Inference is performed on a Dell Latitude 7,420 machine with 16GB of RAM and an 11th generation Intel Core i5 processor with 8 cores. Figure 3.6 (left panel) shows the time taken to fit the one-, two-, and three-step SCMEVMs with graphical and saturated covariance structures.

Table 3.1: Median (2.5% and 97.5% quantiles) bias in the fitted maximum log-likelihood values. Bold values denote the least biased stepwise inference procedure for each covariance structure type and conditioning variable.

Covariance Structure		Independent			Graphical			Saturated		
Number of Excesses	Conditioning Variable	One-step	Two-step	Three-step	One-step	Two-step	Three-step	One-step	Two-step	Three-step
250	1	**-4.8 (-18.3, 8.7)**	-6.2 (-19.3, 8.2)	-6.2 (-19.3, 8.2)	14.0 (7.6, 22.5)	12.8 (6.3, 21.5)	**12.7 (6.7, 21.4)**	14.8 (7.6, 25.1)	13.6 (6.4, 24.1)	**13.5 (7.0, 24.0)**
	2	**-4.2 (-20.6, 11.7)**	-4.9 (-18.9, 10.0)	-4.9 (-18.9, 10.0)	14.2 (7.2, 23.6)	**11.9 (5.4, 21.8)**	**11.9 (5.4, 21.7)**	15.3 (8.0, 25.1)	13.2 (6.4, 23.0)	**13.1 (6.2, 22.9)**
	3	**0.1 (-11.9, 13.1)**	-1.6 (-12.3, 10.7)	-1.6 (-12.3, 10.7)	13.2 (7.1, 22.7)	**11.1 (4.7, 21.3)**	**11.1 (4.8, 21.2)**	15.1 (7.8, 25.6)	**13.2 (5.4, 24.3)**	13.4 (5.6, 24.2)
	4	**-3.7 (-18.2, 11.5)**	-5.2 (-19.4, 9.9)	-5.2 (-19.4, 9.9)	13.6 (8.2, 21.5)	12.3 (5.9, 20.0)	**12.2 (5.9, 19.9)**	14.5 (7.7, 23.1)	**13.3 (6.6, 21.6)**	**13.3 (6.7, 21.6)**
	5	**-13.3 (-34.1, 3.3)**	-14.8 (-30.0, 2.3)	-14.8 (-30.0, 2.3)	13.5 (7.6, 22.2)	12.2 (4.7, 21.3)	**12.0 (5.3, 21.0)**	14.4 (8.7, 24.1)	**12.9 (5.2, 22.5)**	**12.9 (6.2, 22.4)**
500	1	**-20.9 (-35.9, -4.9)**	-22.4 (-39.0, -6.0)	-22.4 (-39.0, -6.0)	14.4 (9.0, 21.3)	**12.8 (7.1, 20.1)**	**12.8 (7.1, 20.0)**	15.2 (9.1, 22.7)	**13.6 (7.6, 21.0)**	**13.6 (7.5, 20.9)**
	2	**-20.1 (-36.5, -0.3)**	-22.0 (-37.0, -2.1)	-22.0 (-37.0, -2.1)	13.8 (7.4, 22.0)	12.0 (4.8, 20.7)	**11.9 (5.4, 20.6)**	14.9 (8.4, 23.4)	13.4 (5.9, 22.4)	**13.3 (6.0, 22.3)**
	3	**-13.4 (-30.9, 2.6)**	-16.2 (-31.5, 1.3)	-16.2 (-31.5, 1.3)	13.0 (7.1, 20.9)	**10.5 (3.4, 19.3)**	10.6 (3.6, 19.3)	14.9 (8.9, 22.9)	12.9 (5.4, 20.8)	**12.8 (5.2, 20.7)**
	4	**-17.9 (-33.9, -0.7)**	-19.9 (-36.2, -3.8)	-19.9 (-36.2, -3.8)	14.2 (8.2, 24.3)	**12.1 (5.0, 21.3)**	**12.1 (5.0, 21.3)**	15.1 (8.5, 24.8)	**13.2 (5.4, 22.1)**	**13.2 (5.4, 22.0)**
	5	**-37.0 (-58.1, -18.8)**	-39.0 (-60.3, -20.7)	-39.0 (-60.3, -20.7)	13.8 (7.5, 21.9)	**12.0 (5.1, 20.5)**	**12.0 (5.1, 20.4)**	14.7 (8.6, 24.1)	**12.9 (6.3, 21.7)**	13.0 (6.4, 21.6)

Table 3.2: Comparison of the average time (seconds) to complete each step of the three-step model fitting procedure across different dimensions.

	Dimension				
Inference step	100	200	300	400	500
Dependence parameters	1.34	2.07	2.84	3.65	4.42
AGG parameters	0.94	1.71	2.65	3.64	4.61
Graphical covariance structure	0.10	0.81	3.61	11.06	30.25
Saturated covariance structure	0.03	0.10	0.22	0.38	0.60

The one- and two-step methods are considerably slower as they require joint maximisation of likelihood functions over parameter spaces with a minimum dimension per conditioning variable of $6(d-1)$ (one-step) and $4(d-1)$ (two-step). Additionally, the higher the dimension of the parameter space, the harder it is to find initial values for the numerical optimisation. The three-step method is more efficient, with considerable time savings when n, d, or both are large.

Figure 3.6 (left panel) suggests that for the three-step method, it may be faster to use the saturated covariance than its sparse graphical counterpart. We repeat the study for these two cases with $n = 4 \times 10^3$ and $d \in \{100, 200, 300, 400, 500\}$, setting the proportion of edges in each graph to be 10% of the maximum possible number of edges. Table 3.2 shows the average time taken to complete each inference step. As expected, the high cost is due to estimation of the graphical structure via the graphical lasso (Friedman et al., 2007). In contrast, the saturated covariance is estimated empirically, i.e., no numerical maximisation is required,

and the computational cost from inverting a $(d-1)$-dimensional correlation matrix is lower. For the saturated model, the inversion could be bypassed if only the correlation matrix were required. However, to be consistent with the other stepwise inference procedures, we use the precision matrix to fit the model, thus inversion of the $(d-1)$-dimensional correlation matrix is required in our case.

3.4.2 Graphical selection

We now replicate the simulation study of Engelke and Hitz (2020, Section 5.5) to assess how well the SCMEVM identifies a specific graphical structure. In this study, $d = 16$ and the data generating mechanism is the Hüsler-Reiss distribution with dependence structure determined by the graph $\mathcal{G}$ shown in Figure 3.3 (left panel). The parameters for each of the $p = 18$ edges are sampled independently from a uniform distribution on $(0.5, 1)$, subject to the constraint that the parameter matrix must be conditionally negative definite on cliques of size three. We take 100 simulated datasets, each of size 10^3.

Algorithm 3.5 is used to infer the optimal graphical structure. We take a majority rule proportion p of 0.5. Setting the dependence thresholds u_{Y_i} to the 0.90-quantile of the standard Laplace distribution results in approximately 100 observations for inferring the graph. Figure 3.3 (right panel) shows a weighted graph with line width and darkness proportional to the number of times the edge is selected across the 100 datasets. After pruning this graph using the majority rule, the true form of $\mathcal{G}$ is clearly identified.

For the graphical lasso step we set $\boldsymbol{\lambda} = \{0.4, 0.41, \ldots, 0.8\}$. Across the simulated datasets, all inferred values of the penalty parameter lay between 0.61 and 0.73. The number of edges, equivalently the density of the dependence structure, was slightly overestimated; graphs with $\{18, 19, 20, 21, 22, 23, 24\}$ edges were selected for $\{38, 29, 14, 8, 7, 2, 2\}$ out of the 100 datasets. The results are not overly sensitive to the dependence threshold, with similar findings when using the 0.80- or $0.95-$quantiles. Sensitivity to the majority rule proportion p is also minimal (see Supplementary Material).

Despite using different underlying models and having different graphical selection processes, both our method and the Engelke and Hitz (2020) method accurately infer the underlying graphical structure. Not unexpectedly, the Engelke and Hitz (2020) method performs slightly better, either because their model is also the data generating mechanism Engelke and Hitz (2020) or because the data are AD. In contrast, AD is an edge case in the SCMEVM. We could more accurately capture AD data by setting $\alpha_{j|i} = 1$ and $\beta_{j|i} = 0$, but in practice, it is usually preferable to retain the flexibility to capture both AD and AI. A second simulation

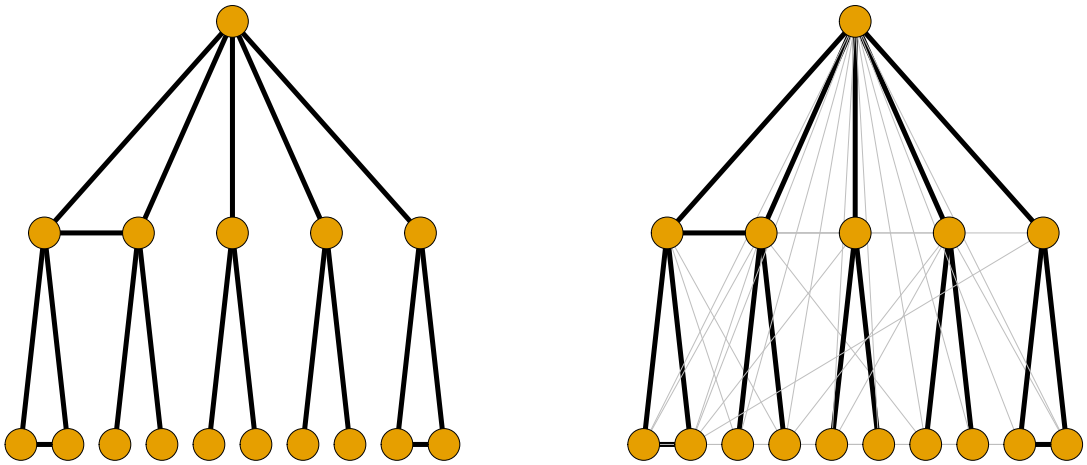

Figure 3.3: True underlying graphical structure (left) and the inferred graphical structure (right), with line width and darkness indicating the number of times each edge was selected across 100 samples.

study, with data generated from the asymptotically independent MVG, is provided in the Supplementary Material. The results there demonstrate that our method is more generally applicable than that of Engelke and Hitz (2020).

3.4.3 Mixture data

To test the flexibility of our model, we consider data with a mixture of extremal behaviour; comparable studies for either full AI or AD can be found in the Supplementary Material. The mixture data is sampled as follows. Firstly, (X_1, X_2, X_3) are sampled from a multivariate Pareto distribution with a fully connected graph and transformed to standard Gaussian margins. Secondly, $(X_4, X_5) \mid X_3 = x_3$ are sampled from a MVG distribution with a fully connected graph. Then $\boldsymbol{X} = (X_1, \ldots, X_5)$ has dependence structure consistent with $\mathcal{G}$ in Section 3.4.1. A total of 200 datasets are simulated. Note, while such a data structure may not be realistic in practice, this example is helpful to assess the range of behaviour that can be captured by the model.

Each of the one-(Algorithm 3.1), two-(Algorithm 3.3), and three-step (Algorithm 3.4) procedures are used to fit the SCMEVM with graphical covariance. The SCMEVMs with independent and saturated covariances are only implemented for the three-step procedure. For comparison with existing methods, we fit the CMEVM as described in Heffernan and

Figure 3.4: Boxplots of empirical and model-based estimates of $\Gamma_{|i}$, for each $i \in V$, when the data is generated from a mixture distribution. Each row corresponds to the conditioning variable i, and each column corresponds to the correlation parameter. The colour of the boxplots distinguishes the different models. The black dashed line indicates the $y = 0$ line.

Tawn (2004) and the graphical extremes model of Engelke and Hitz (2020) (EHM). For the SCMEVMs with graphical covariance structure and the EHM, we use the graph $\mathcal{G}$ defined in Section 3.4.1.

Selection of the threshold u_{Y_i} requires some consideration since subsets of the components may be fully AD, fully AI, or a mixture of both. Consequently, the rate of convergence to the limiting dependence structure varies by conditioning variable. Theoretically, it would be preferable to use a different threshold for each conditioning variable. However, since convergence rates are unknown in practice, we proceed by setting the threshold for each component to be the 0.90-quantile of the standard Laplace distribution, giving approximately 500 excesses per conditioning variable.

Figure 3.4 shows empirical and model-based estimates of the conditional precision matrix $\Gamma_{|i}$. The empirical estimates are obtained by inverting the empirical correlation matrix of $\boldsymbol{Y} \mid Y_i > u_{Y_i}$ and then excluding the ith row and column. We first compare the precision

matrix estimates across models; the estimates are almost identical for the graphical and saturated SCMEVMs, confirming there is negligible loss in using the former. The magnitudes of the non-zero entries in the residual precision matrices cannot be directly compared with their empirical equivalents due to the non-linear transformation (3.2.1). We can compare the location of the zero entries and find that the graphical and saturated SCMEVM estimates are consistent with their empirical equivalents. This suggests that the graphical structure of the residual distribution is inherited from the underlying multivariate distribution, a result consistent with the theoretical findings of Casey and Papastathopoulos (2023) and similar simulation studies for cases with full AI or full AD (see Supplementary Material).

Figure 3.5 shows the bias in the estimates for the tail probabilities $p_1 = \mathbb{P}[X_1 > v_1, X_2 > v_2 \mid X_3 > u_3]$ and $p_2 = \mathbb{P}[X_3 > v_3, X_4 > v_4 \mid X_5 > u_5]$, where u_i and v_i are the 0.90-quantile and 0.95-quantile of X_i, respectively. Note that the quantiles and "true" probabilities are obtained from a sample of size 10^6 from the underlying generating mechanism; their their actual values could be derived by disentangling the complex conditioning in the underlying generating mechanism. Estimates from the CMEVM and the SCMEVM with graphical or saturated covariances are unbiased for p_1 and exhibit a small positive bias for p_2. The EHM estimates of p_1 are similarly unbiased. However, the estimates of p_2 are considerably more positively biased compared to (S)CMEVM models. These patterns are consistent across estimates of the remaining 73 conditional probabilities of the form $\mathbb{P}[\boldsymbol{X}_A > v \mid X_i > v]$ for $A \subseteq V_{|i}$ and $i \in V$. When conditioning on sites 4 and 5, estimates from the CMEVM and SCMEVM always exhibit some small positive bias. Since the bias is consistent across the two models, we postulate that the cause is the rate at which the different components of the generating distribution converge to the tail distribution. For instance, the Pareto (Gaussian) distribution converges quickly (slowly) to its limiting distribution. Thus, conditioning on a component with quick/slow convergence will likely impact the rate of convergence of the dependent component. Since components 4 and 5 are (conditionally) Gaussian, their convergence will be slow, impacting the subsequent probability estimates if the model has not fully converged. The bias may be reduced if lower dependence thresholds were employed for these sites.

Finally, we consider the mean absolute error (MAE) and root mean square error (RMSE) of the estimates for all 75 conditional tail probabilities. The SCMEVM outperforms both the EHM and CMEVM, minimising the MAE for 73% of the probabilities and the RMSE for 76% of them. In contrast, the CMEVM (EHM) minimises the MAE 15% (12%) and the RMSE 12% (12%) of the time. This suggests that even for moderate d, the SCMEVM has greater predictive precision and accuracy than competitor methods.

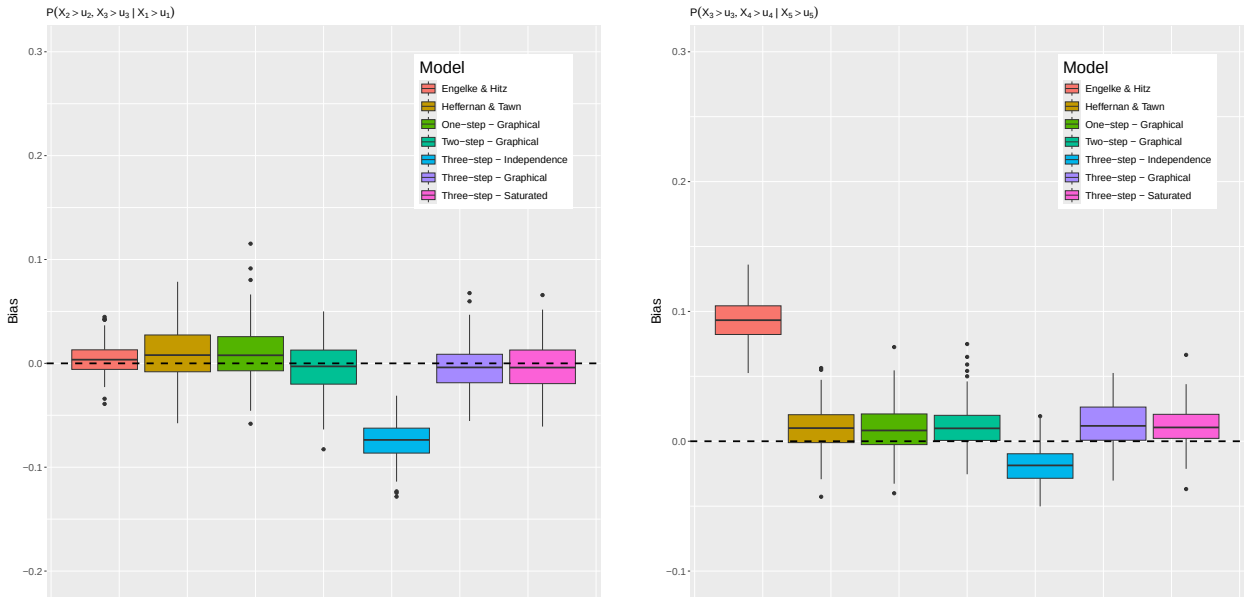

Figure 3.5: Boxplots of the bias in $p_1 = \mathbb{P}[X_1 > v_1, X_2 > v_2 \mid X_3 > u_3]$ (left) and $p_2 = \mathbb{P}[X_3 > v_3, X_4 > v_4 \mid X_5 > u_5]$ (right). The fill of the boxplots distinguishes the different models. The black dashed line indicates the $y = 0$ line.

3.5 Application

We apply our model to discharge data from the upper Danube River basin described in Section 3.1. For the margins, we use the empirical-generalised Pareto distribution model from Section 3.3. The dependence threshold u_{Y_i} used is the 0.80-quantile of the standard Laplace distribution for all $i \in V$, resulting in around 85 excesses per station.

We use the three-step procedure to fit the SCMEVM with graphical covariance structure given by the undirected tree induced by the flow connections of the river (Figure 3.1, left panel); we also fit the EHM with the same structure. Since the flow connection tree may not be the optimal structure for describing the dependence structure of the extremes, we also apply the three-step procedure to fit both the SCMEVM with saturated covariance and the SCMEVM with an inferred graphical structure. The inferred graphical structure, obtained from Algorithm 3.5 is shown in Figure 3.6 (right panel). This graph has 127 edges compared to the maximum possible 465 edges. Lastly, for prediction, we simulate datasets from the fitted models using Algorithm 3.6 with $N = 20n$ and $u = u_{Y_i}$.

Bootstrapped estimates of the coefficient of tail dependence $\eta_{i,j}(u)$ for $i, j \in V$, $i > j$, and $u \in \{0.8, 0.85, 0.9\}$, are obtained using 200 bootstrapped samples of the data. For each bootstrapped dataset, both empirical and model-based estimates of $\eta_{i,j}(u)$ are obtained. The point estimates in Figure 3.7 are the medians of the two sets of estimates. The SCMEVMs

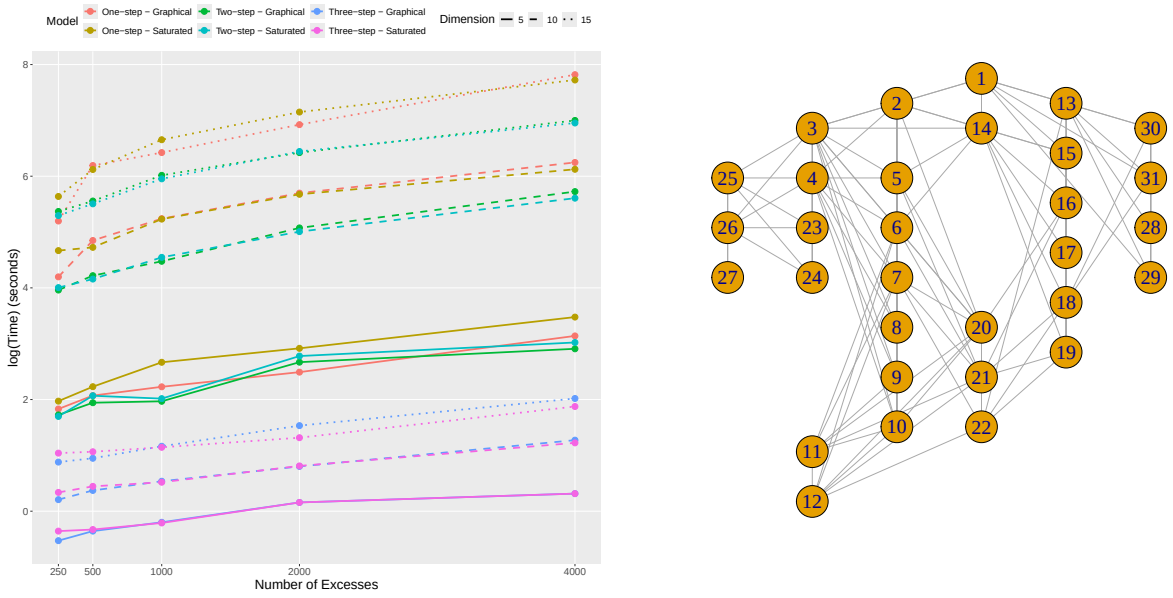

Figure 3.6: Timing comparison (log scale) of the SCMEVMs (left) for various sample sizes, dimensions, denoted by the line type, and models, denoted by the line colour. Inferred graphical structure of the upper Danube River basin using Algorithm 3.5 (right).

describe the empirical dependence better than the EHM for both flow-connected (triangles) and flow-unconnected (circles) stations, and across all values of u. This highlights the value of a model that captures a range of extremal dependence classes. As noted, numerous extensions to the EHM have been proposed (Engelke et al., 2024a) that allow for any sparse graphical structure. Therefore, we compare both the learnt graphical structure and the predictive performance of the learnt model, for the SCMEVM and EGlearn (Engelke et al., 2025) in the Supplementary Material.

Figure 3.7 shows that all pairs of stations appear to exhibit AI with positive association, $\eta(u) \in (0.5, 1)$, while analogous plots of $\chi(u)$ (see Supplementary Material) imply all pairs have AD. However, for stations with weaker (stronger) extremal dependence, the two measures decrease (stay close to 1) as u increases. This supports the plausible conclusion that some pairs of stations, particularly those that are flow-unconnected, exhibit AI while others, particularly those that are flow-connected, exhibit AD.

For $u = 0.8$, the SCMEVM with saturated covariance performs noticeably better than the graphical SCMEVM with structure given by the undirected tree induced by the flow connections. This suggests that the extremal dependence is influenced by factors beyond the river structure (see also Asadi et al. (2015)). To support this hypothesis, Figure 3.8 shows the difference in the empirical and model-based estimates of $\eta_{i,j}(0.8)$ for each pair $i, j \in V$, where the model-based estimates are from the SCMEVM with graphical covariance using

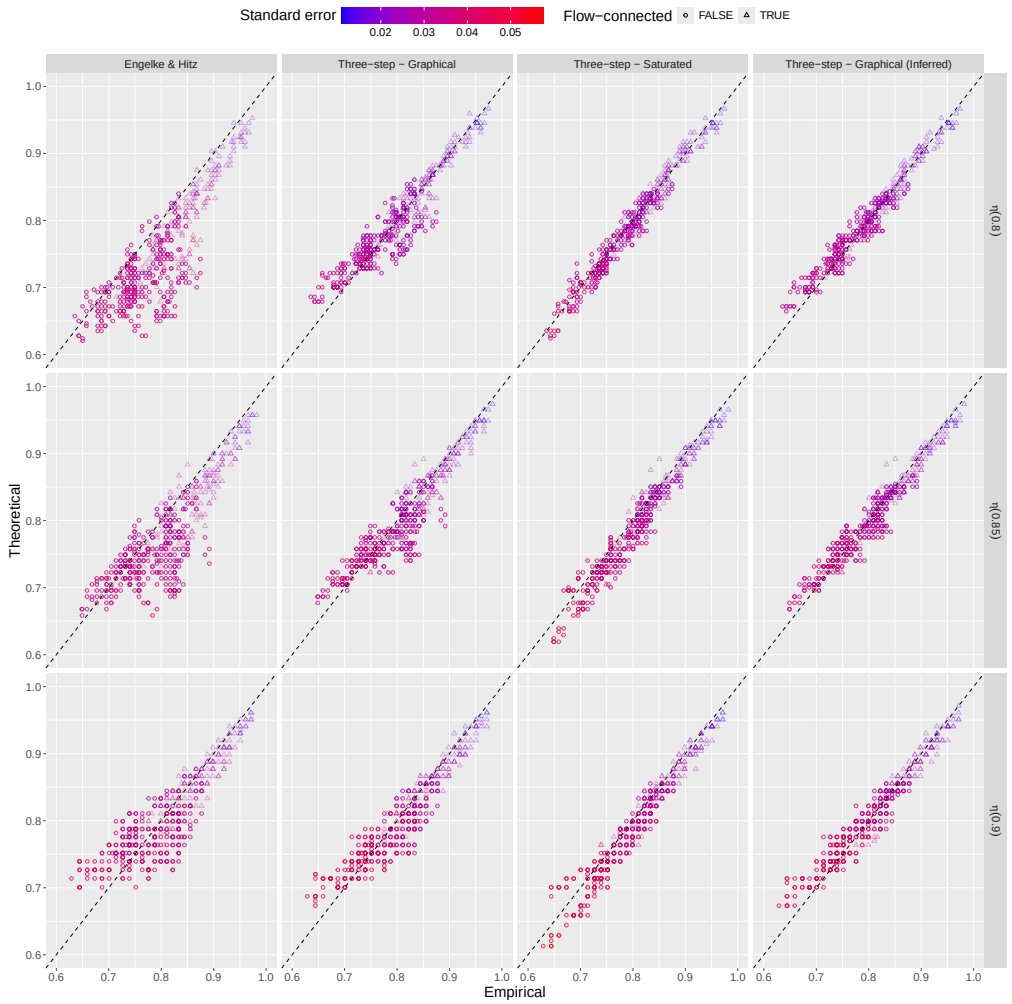

Figure 3.7: Empirical and model-based estimates of $\eta_{i,j}(u)$ for $u \in \{0.8, 0.85, 0.9\}$ (top to bottom), and $i, j \in V$ but $i > j$. Model-based estimates use the EHM (an AD based model) (left) and the three-step SCMEVM (an AI based model) with graphical covariance (centre left), both with structure given in Figure 3.1 (left panel), the three-step SCMEVM with saturated covariance (centre right) and graphical covariance (right) with structure given in Figure 3.6 (right panel). Black dashed lines show $y = x$. Circles (triangles) show flow-connected (flow-unconnected). The colour shows the standard error of the model-based estimates.

the induced undirected tree. We observe that underestimation predominantly occurs for flow-unconnected stations. For example, the dependence between stations 11-12 and 16-22 is considerably underestimated. While these two sets of stations are flow-unconnected, the sources of their tributaries are geographically close and at similar altitudes, thus, stronger dependence than suggested by the lack of flow connection is not unexpected. Similar observations are made when comparing the Isar (stations 14 - 19) and Salzach (stations 28 - 31) tributaries, as well as stations 23 - 24 and 25 - 27. Furthermore, the inferred graphical

structure in Figure 3.6 (right panel) shows many connections between sites on geographically neighbouring tributaries. Indeed, using the inferred graphical structure in the SCMEVM drastically improves the model fit (Figure 3.7, right panel), and resolves the systematic underestimation caused by using a saturated covariance structure.

Returning to Figure 3.7, we observe that the SCMEVMs underestimate dependence at higher thresholds, particularly for flow-unconnected stations with weaker associations. In contrast, the EHM becomes less biased as the threshold increases, although its bias at lower thresholds is larger than the bias in the SCMEVMs at higher thresholds. Moreover, the cross-site variability in the model-based estimates is higher for the EHM than the SCMEVMs, regardless of level, particularly for stations with weaker associations. In conclusion, the SCMEVMs more accurately and consistently represent the extremal dependence in the data than the EHM.

3.6 Discussion

In this paper, we have extended the conditional multivariate extreme value model (Heffernan and Tawn, 2004) by replacing the non-parametric residual distribution with a flexible, fully parametric model. This overcomes the curse of dimensionality that arises when extrapolating from a non-parametric estimate of a high-dimensional distribution, resulting in more accurate and reliable predictions in high dimensions. Our proposed parametric model is the MVAGG distribution. The copula-based construction of the MVAGG, which combines asymmetric generalised Gaussian margins with a multivariate Gaussian dependence structure, facilitates efficient statistical inference, as the margins and dependence structures can be inferred separately in a stepwise manner. Further, separate estimation of the marginal and dependence parameters for the MVAGG is computationally efficient and loses no information compared to joint estimation of all parameters.

To reduce the parameter space, we propose using a graphical structure to induce sparsity into the precision matrix of the MVG copula. In addition to reducing the number of unknown parameters to be estimated, this provides a mechanism to infer the dependence structure if it is not already known. Despite the sparsity induced by the graphical structure, model fitting is substantially more expensive than using a saturated covariance structure due to the required numerical optimisation. Therefore, while graphical structures can be learnt and implemented, the model in its current form may not be suitable for very large dimensions and further work is needed to address this computational hurdle.

Our analysis of the upper Danube River basin dataset suggests the SCMEVM captures the

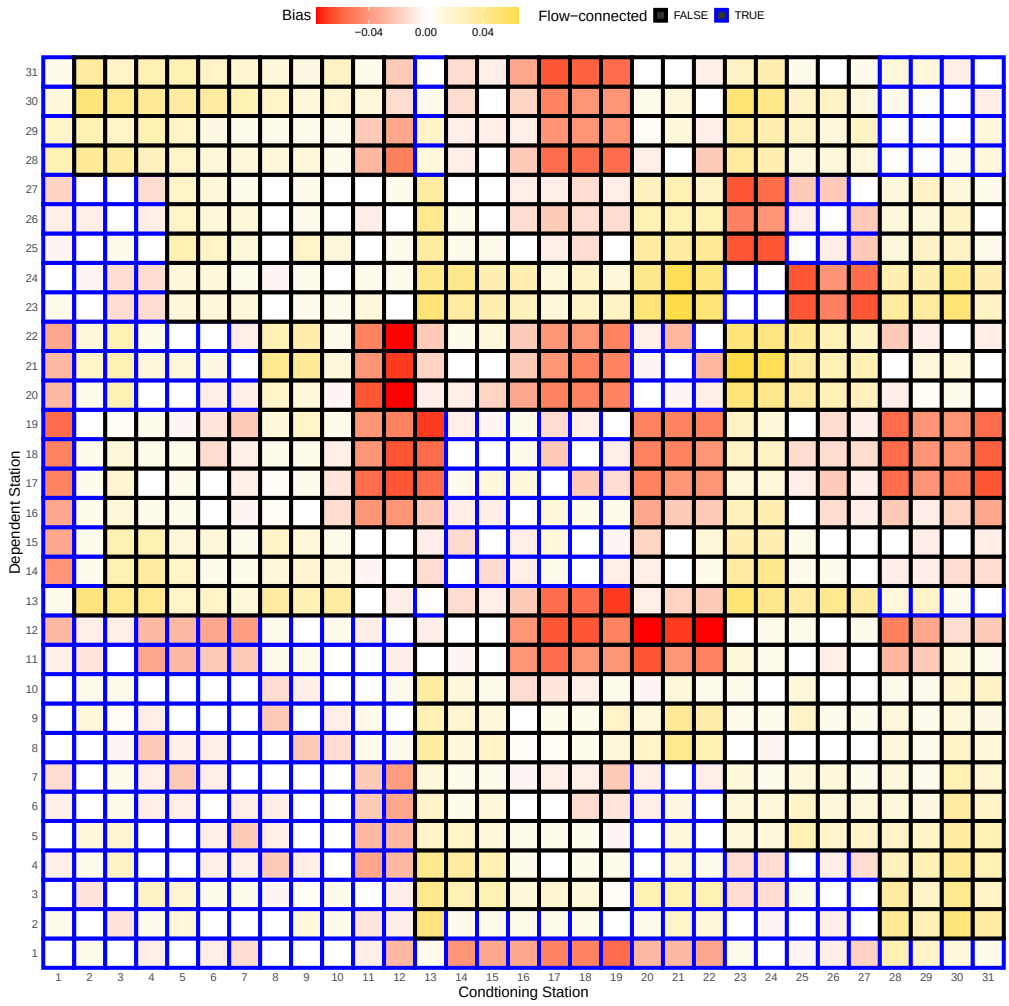

Figure 3.8: Difference between empirical and median of model-based estimates of $\eta_{i,j}(0.8)$ for each $i, j \in V$ for the SCMEVM with a graphical covariance, where the graphical structure is assumed to be the undirected tree induced by the flow connections in Figure 3.1 (left panel). Under- and over-estimation from the model is represented by red and gold squares, respectively. Flow-connected and flow-unconnected stations are represented by blue and black borders, respectively.

dependence between stations more effectively than the graphical extremes model of Engelke and Hitz (2020), highlighting the benefit of a model that can capture both AD and AI. Furthermore, the SCMEVM, based on the undirected tree inferred from the flow connections of the river network, does not perform as well as either the graphical structure inferred from the graphical lasso or the saturated covariance matrix, highlighting the complex dependence structure in the data that is not solely captured by the data's underlying graphical structure. Thus, a possible alternative that incorporates the river network structure would be to add a second covariance matrix based on Euclidean distance (Asadi et al., 2015) into the MVG copula component of the SCMEVM.

Finally, as with the graphical extremes model of Engelke and Hitz (2020), our model only allows predictions at measured locations. Parameterising the Gaussian copula kernel with a Matérn or Whittle-Matérn correlation function (Bolin et al., 2024), where distance is measured along the graphical structure, would allow extrapolation to unobserved locations. The generally strong correspondence between the empirical and model-based estimates of η for the flow-connected sites from the SCMEVM provides confidence that such a model would result in reliable extrapolations.

Supplementary Material to "Conditional Extremes with Graphical Models"

S3.1 Prediction from the conditional multivariate extreme value model

The original conditional multivariate extreme value model (CMEVM) uses a semi-parametric algorithm for prediction to avoid over-reliance on the working distributional assumptions used for parameter estimation (see Section 2.1 of the main text). Specifically, prediction is performed by non-parametrically sampling with replacement from the empirical distribution of the fitted residuals. Such sampling suffers from the curse of the dimensionality (Nagler and Czado, 2016), meaning the predictive performance of the CMEVM decreases as the dimension increase. To demonstrate this, we perform a simple simulation study and compare the predictive performance of the CMEVM to the structured CMEVM (SCMEVM) proposed in Section 2 of the main text.

Consider a simple undirected graph $\mathcal{G} = (V, E)$ with vertex set $V = \{1, \ldots, d\}$ and edge set $E \subseteq \{\{j, k\} \mid j, k \in V, j \neq k\}$. We set $d = 20$ and randomly select the edges in the graph such that the number of edges is approximately 20% of the number of edges in the full graph. We simulate 200 datasets of size 250 for $\boldsymbol{Y} \mid Y_i > u_{Y_i}$ as per Section 4.1 of the main text i.e., we simulate $Y_i | Y_i > u_{Y_i}$ from a standard exponential distribution and obtain $\boldsymbol{Y}_{|i} \mid Y_i > u_{Y_i}$ using equation (2.3) of the main text with $\boldsymbol{Z}_{|i}$ simulated from a multivariate asymmetric generalised Gaussian (MVAGG) distribution (Section 2.2 of the main text). We set the dependence threshold u_{Y_i} to the 0.8-quantile of the standard Laplace distribution. True dependence and asymmetric generalised Gaussian (AGG) parameters are independently sampled from uniform distributions on $(0.1, 0.3)$ for α_j, $(0.1, 0.2)$ for β_j, $(-1, 1)$ for ν_j, $(0.5, 1)$ for κ_{1_j}, $(1.5, 2)$ for κ_{2_j}, and $(0.8, 2.5)$ for δ_j, for each $j \in V$.

For computational purposes, we consider a single conditioning component i selected at random from V; similar results can be obtained when conditioning on different components. Predictive performance is assessed on Laplace margins only since the probability integral transform (PIT) used to back-transform to the original margins does not alter the dependence structure. For each data set, we fit the (i) CMEVM, (ii) three-step SCMEVM with graphical covariance structure, and (iii) three-step SCMEVM with saturated covariance. For (ii), the graph is assumed to be known and correctly specified above. For prediction, we used datasets of size 5×10^6 for $\boldsymbol{Y} \mid Y_i > u_{Y_i}$ simulated from the fitted models using the methods described in Section 4.4 of Heffernan and Tawn (2004) for (i) and Algorithm 1 in Section 5.1

of Wadsworth and Tawn (2022) for (ii) and (iii).

Figure S3.1 shows $\mathbb{P}[\boldsymbol{Y}_A > v \mid Y_i > v]$, on the exponential scale, for 500 different sets $A \subseteq V_{|i} = V \backslash \{i\}$ such that $|A| = 3$; the sets were chosen at random. We set v to be the 0.999-quantile of the standard Laplace distribution, which would approximately correspond to a 1 in 10 year event if we had daily data. The truth is obtained empirically using a single sample of size 10^7 from the true distribution, while the model-based estimates are the median of the model-based point estimates from each of the 200 samples. The CMEVM consistently underestimates the probabilities, whereas the SCMEVMs perform much better, particularly for those probabilities that are very close to 0. The standard error of the model-based estimates (on the original scale) appear to be lower for the SCMEVMs compared to the CMEVM. The SCMEVMs do slightly underestimate the probabilities, however, we anticipate this could be resolved by increasing the size of the prediction datasets.

The CMEVM underestimates the probabilities because it allows neither interpolation nor extrapolation of the fitted residuals, resulting in "rays" in data simulated from the fitted model. This can be seen in Figure S3.2, which shows $2,000$ randomly selected points from data simulated from the fitted models for (i) and (ii). Data used to fit the models is also shown. The CMEVM does not accurately capture dependence between components 2 and 13, but the SCMEVM does much better. This pattern will only be exacerbated as the dimension increases. Therefore, the predictive power of the CMEVM will diminish as (1) v approaches the upper end-point of the distribution, (2) the size of set A increases, and (3) the dimension of the problem increases. The SCMEVM overcomes such limitations by using a flexible, fully parametric distribution for the residuals.

S3.2 Marginal distributions for the residuals

In Section 2.2 of the main text, we propose an adaptation of the generalised Gaussian model used by Wadsworth and Tawn (2022). Here we show that the proposed alternative, the asymmetric generalised Gaussian distribution, improves the overall model fit. For $d = 20$ and a graph $\mathcal{G} = (V, E)$ with edges randomly selected such that the number of edges is approximately 20% of the number of edges in the full graph, we simulate 200 datasets of size 5000 from a multivariate Gaussian (MVG) distribution with mean vector $\boldsymbol{\mu}$ and correlation matrix Σ. The components μ_j are independently sampled from a uniform distribution on $(-5, 5)$, and the correlation matrix is associated with $\mathcal{G}$. For each replicate $\boldsymbol{X}$, we transform the margins onto standard Laplace margins $\boldsymbol{Y}$, as per Section 3.1 of the main text, before fitting the SCMEVM. We set the dependence thresholds u_{Y_i} to the 0.90-quantile of the

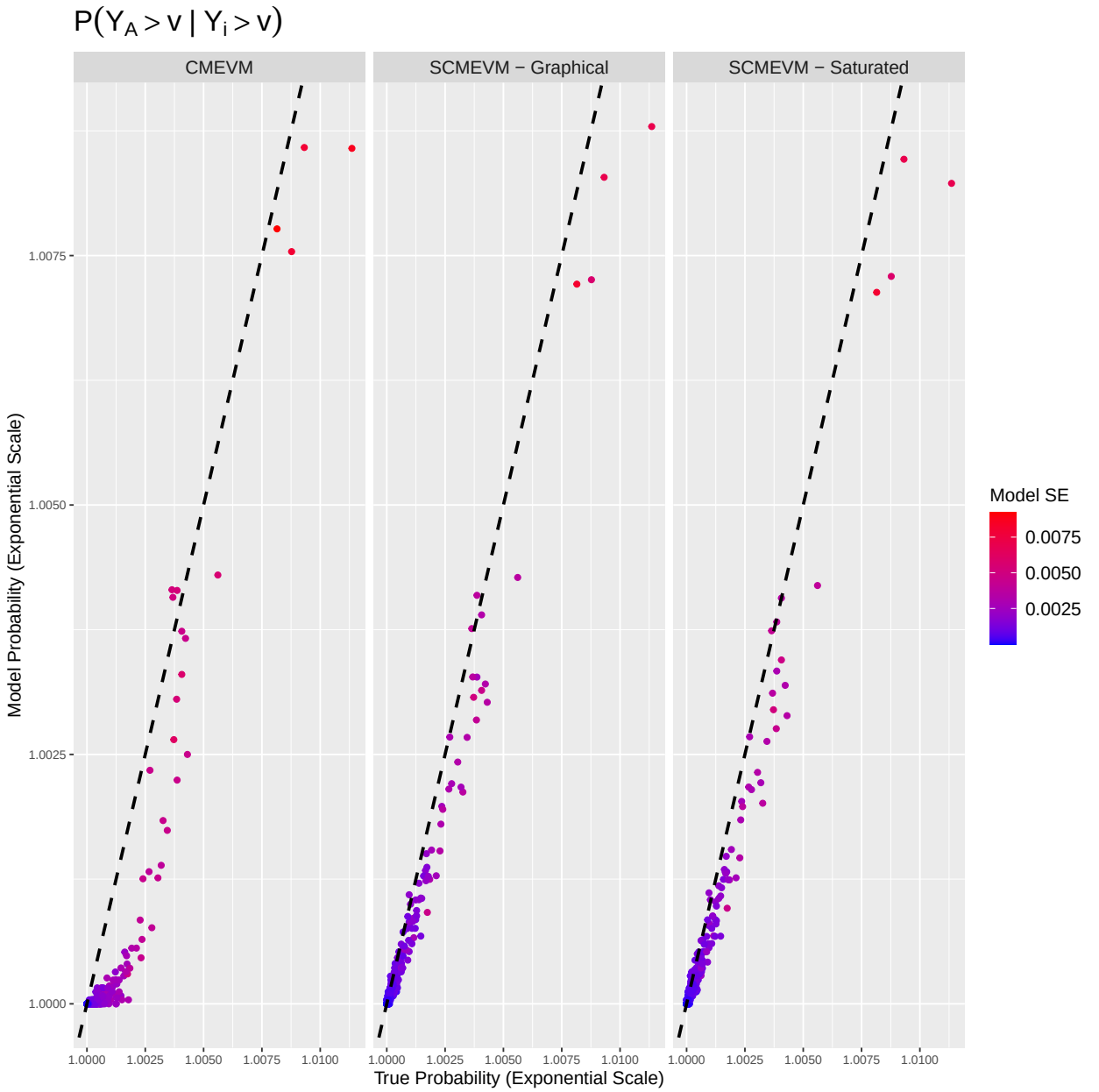

Figure S3.1: Empirical and model-based estimates of $\mathbb{P}[\boldsymbol{Y}_A > v \mid Y_i > v]$ for a randomly selected component $i \in V$, 500 randomly selected sets $A \subseteq V_{\setminus i}$ such that $|A| = 3$, and v is the 0.999-quantile of the standard Laplace distribution. Model-based estimates use the CMEVM (left), the three-step SCMEVM with graphical covariance (centre) with structure described in Section S3.1, and the three-step SCMEVM with saturated covariance (right). The colour shows the standard error of the model-based estimates. Black dashed lines show the $y = x$.

standard Laplace distribution.

To fit the SCMEVM, we use the three-step (Algorithm 3.4 of the main text) procedure and assume a saturated covariance structure. For comparison, we fit the same model but with

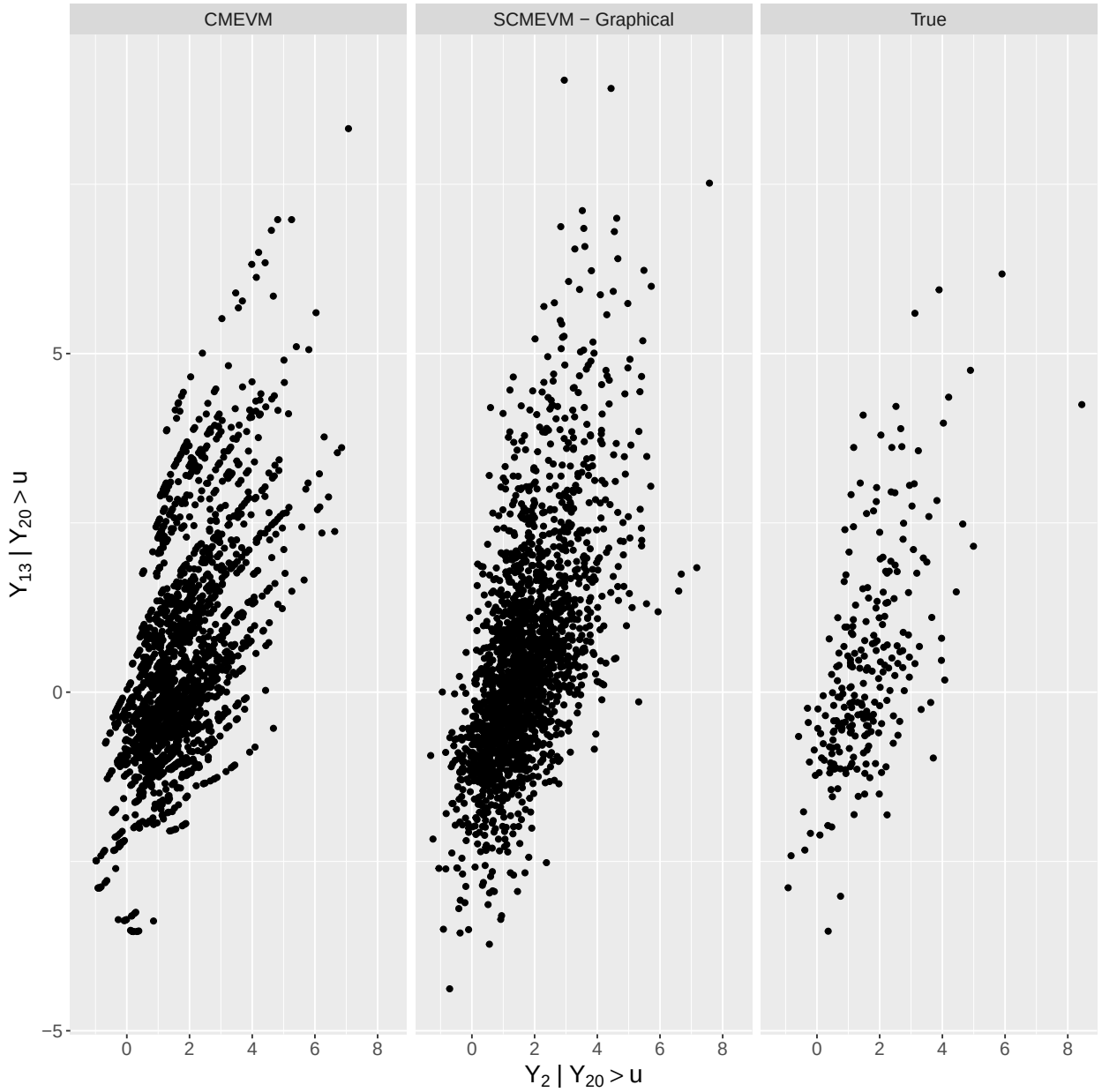

Figure S3.2: Scatter plots for Y_2 and Y_{13} given $Y_{20} > u_{Y_{20}}$. The points correspond to $2,000$ randomly selected data points from a sample of size 5×10^6 simulated from the fitted model for the CMEVM (left), and the three-step SCMEVM with graphical covariance (centre) with structure described in Section S3.1. Also shown are the 250 points used to fit the models (right).

generalised Gaussian margins for the residual distribution. We then simulate samples of size $N = 20n$ from each of the conditional models (see Section 3.3 of the main text for the simulation algorithm). Figure S3.3 compares the median of the empirical and model-based estimates of $\eta_{i,j}(u)$ for $i, j \in V, i > j$, and $u \in \{0.95, 0.99\}$, over the 200 datasets. The only difference between the left and right panels is the margins used in the residual distribution,

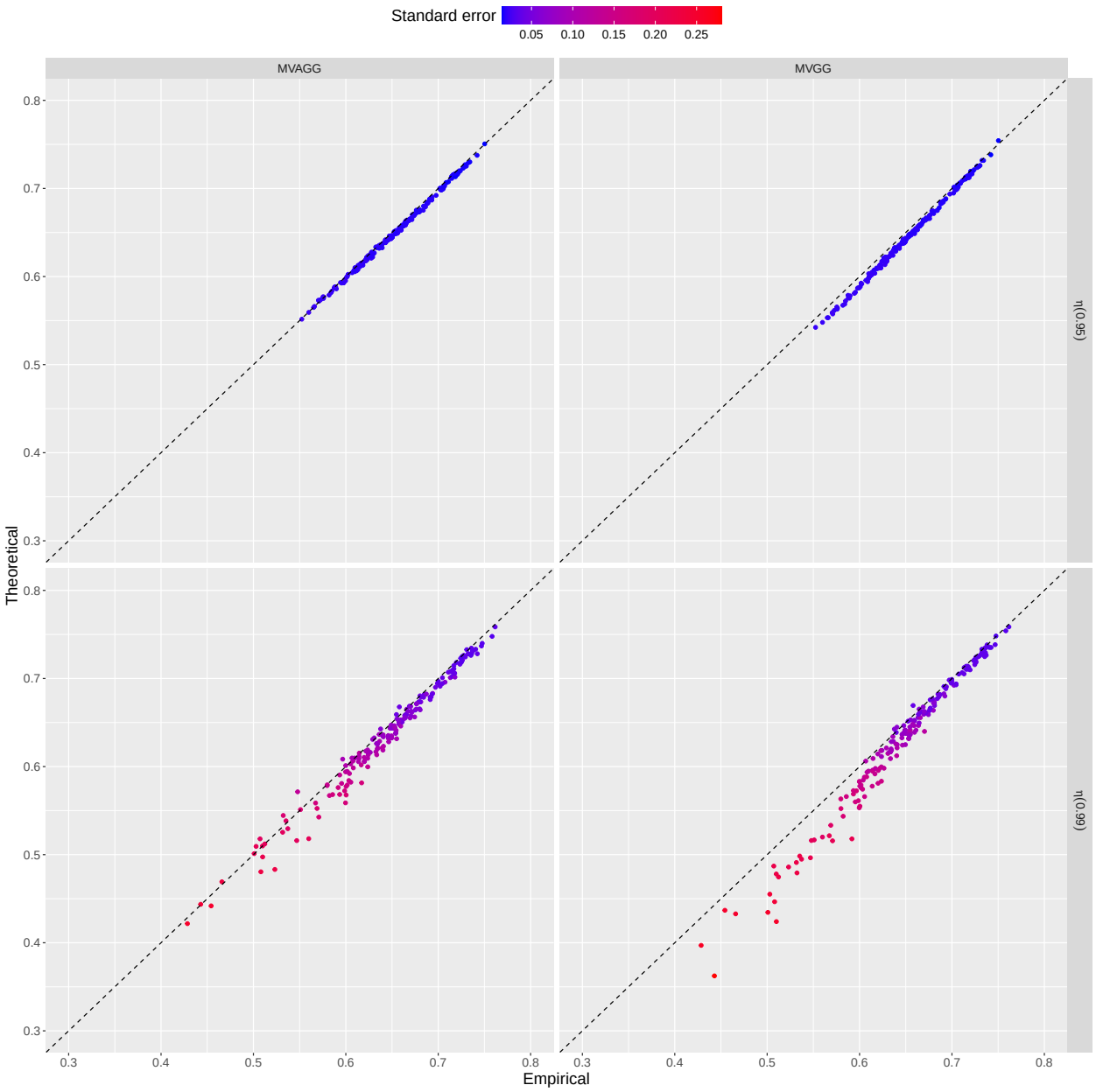

Figure S3.3: Empirical and model-based estimates of $\eta_{i,j}(u)$ for $u \in \{0.95, 0.99\}$ (top to bottom), and $i, j \in V$, but $i > j$. Model-based estimates use the three-step SCMEVM with residuals having a saturated covariance and either asymmetric generalised Gaussian (left) and generalised Gaussian (right) margins. Black dashed lines show $y = x$. The colour shows the standard error of the model-based estimates.

which are AGG and generalised Gaussian, respectively. The generalised Gaussian underestimates the true value of η, particularly for the pairs with weaker dependence. In contrast, the AGG tends to capture η reasonably well with no increase in the standard error despite the additional model parameter. Therefore, the AGG is necessary to obtain accurate predictions from the model.

S3.3 Additional figures and simulation studies for Section 4.1

Section S3.3.1 contains additional figures for the simulation study of Section 4.1 in the main text. Also shown are two additional simulation studies to assess the model performance in the presence of either strong positive (Section S3.3.2) or weak negative (Section S3.3.3) associations. Throughout this section, data are simulated from the SCMEVM with a graphical covariance structure given by $\mathcal{G} = (V, E)$, $V = \{1, \ldots, 5\}$, and $E = \{\{1,2\},\{1,3\},\{2,3\},\{3,4\},\{3,5\},\{4,5\}\}$.

S3.3.1 Weak positive dependence

For this study, recall that the true dependence and AGG parameters were selected at random by sampling from a uniform distribution on $(0.1, 0.5)$ for α_j, $(0.1, 0.3)$ for β_j, $(-5, 5)$ for ν_j, $(0.5, 2)$ for κ_{1_j}, $(1.5, 3)$ for κ_{2_j}, and $(0.8, 2.5)$ for δ_j, for each $j \in V$. Figures S3.4, S3.5 and S3.6 show the bias in $\hat{\boldsymbol{\beta}}_{|i}$, the AGG parameters, and $\hat{\Gamma}_{|i}$ respectively. Similar to the plot for $\hat{\boldsymbol{\alpha}}_{|i}$ in the main text, we omit the maximum likelihood estimates (MLEs) from the stepwise methods in cases where they are, by construction, identical to estimates that are already presented. For $\hat{\Gamma}_{|i}$ we exclude those models that assume independent residuals since these are consistently biased. The findings are very similar to the main text: all models are unbiased across all parameters; the two- and three-step methods show slightly more cross-sample variability in their bias; variability in bias decreases as sample size increases.

S3.3.2 Strong positive dependence

We repeat the simulation study from Section 4.1 of the main text but with strong, positive correlations (> 0.48). The other parameters remain unchanged from Section S3.3.1. Boxplots of the parameter estimates (not included) are almost identical to what was seen with weak positive associations. To compare the three stepwise procedures, we compare the bias in the maximum log-likelihood values, see Table S3.1. The models with independent residuals are biased; this is expected because the dependence structure is clearly misspecified. The bias is lower in the case when we condition on component 3 because this results in exact independence between (W_1, W_2) and (W_4, W_5). This result was not seen in the study from the main text due to the lower correlations used there. Similar to the results shown in the main text, models with a graphical or saturated dependence exhibit a small positive bias, but its magnitude is similar across all stepwise procedures. This supports our claim that the stepwise inference procedures result in no loss of information. Further, the three-step

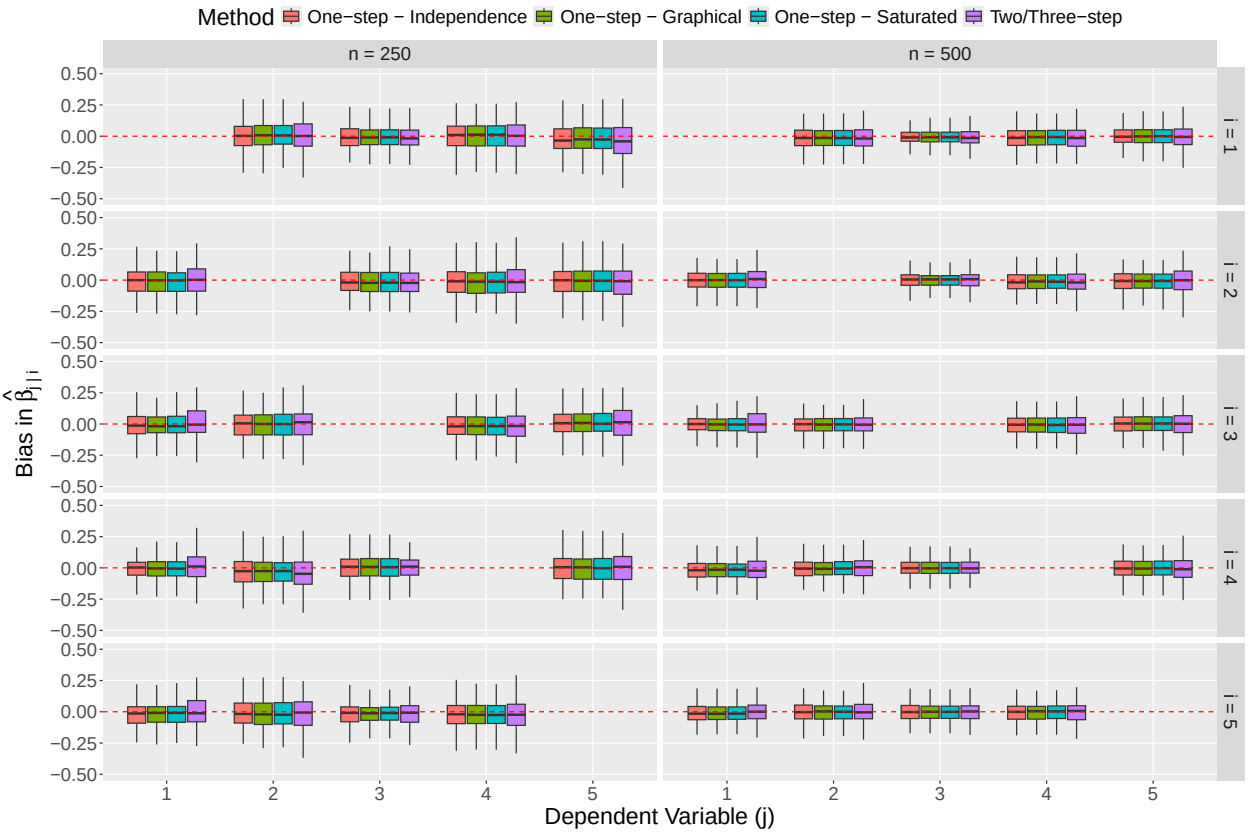

Figure S3.4: Boxplots detailing the bias of $\hat{\beta}_{j|i}$ for distinct $i, j \in V$. Each row corresponds to the conditioning variable i, and each column corresponds to the sample size. The different models are denoted by the fill of the boxplots. Red dashed lines show $y = 0$.

model is least biased when we assume the residuals have a graphical or saturated dependence structure. Finally, the bias for the models with independent residuals increases with sample size, while the bias is of similar magnitude for both sample sizes for models with graphical and saturated covariance structures. This suggests the latter models are more robust to changes in the sample size.

Table S3.1: Median (2.5% and 97.5% quantiles) bias in the fitted maximum log-likelihood values for data from the SCMEVM with strong positive associations. Bold values denote the least biased stepwise inference procedure for each covariance structure type and conditioning variable.

Covariance Structure		Independent			Graphical			Saturated		
Number of Excesses	Conditioning Variable	One-step	Two-step	Three-step	One-step	Two-step	Three-step	One-step	Two-step	Three-step
250	1	**-135.7 (-170.2, -98.9)**	-137.3 (-171.1, -99.6)	-137.3 (-171.1, -99.6)	14.7 (8.0, 22.5)	12.7 (6.1, 20.8)	**11.9 (5.2, 20.6)**	15.6 (8.4, 24.0)	13.5 (6.6, 22.1)	**12.8 (5.9, 21.4)**
	2	**-138.5 (-176.4, -100.0)**	-139.7 (-172.4, -101.3)	-139.7 (-172.4, -101.3)	14.5 (8.0, 23.0)	11.8 (4.7, 21.6)	**11.3 (4.1, 21.0)**	15.4 (9.1, 24.8)	12.8 (5.8, 23.2)	**12.1 (4.9, 22.3)**
	3	**-39.8 (-86.1, -18.7)**	-41.4 (-59.2, -20.4)	-41.4 (-59.2, -20.4)	13.3 (6.7, 22.8)	11.1 (4.3, 20.6)	**11.0 (4.1, 20.3)**	15.1 (7.8, 25.7)	13.2 (4.8, 23.6)	**13.1 (5.2, 23.2)**
	4	**-140.0 (-174.1, -94.0)**	-140.8 (-175.2, -96.7)	-140.8 (-175.3, -96.7)	14.1 (6.9, 22.9)	11.9 (4.7, 19.3)	**11.3 (4.2, 18.7)**	15.2 (8.3, 24.1)	12.9 (5.5, 20.6)	**12.2 (4.4, 20.2)**
	5	**-137.5 (-178.1, -105.5)**	-137.8 (-174.8, -106.5)	-137.8 (-174.8, -106.5)	14.0 (8.0, 22.6)	11.5 (3.2, 20.8)	**10.7 (2.8, 20.5)**	15.1 (8.9, 24.3)	12.5 (4.8, 22.1)	**11.8 (4.4, 21.7)**
500	1	**-280.2 (-326.7, -226.3)**	-281.9 (-327.7, -228.3)	-281.9 (-327.7, -228.3)	13.7 (7.6, 21.1)	11.9 (5.3, 19.2)	**11.2 (4.7, 18.3)**	14.7 (8.7, 22.2)	13.0 (6.2, 20.7)	**12.1 (5.6, 19.9)**
	2	**-286.9 (-332.7, -240.5)**	-289.0 (-333.7, -242.0)	-289.0 (-333.7, -242.0)	13.7 (7.7, 22.8)	11.0 (3.3, 19.5)	**10.5 (2.6, 18.7)**	14.7 (7.6, 24.1)	11.7 (3.8, 21.3)	**11.3 (3.1, 20.3)**
	3	**-95.3 (-126.8, -67.7)**	-97.1 (-123.7, -71.9)	-97.1 (-123.7, -71.9)	12.9 (7.1, 21.3)	10.3 (3.7, 18.7)	**10.1 (3.5, 18.3)**	14.8 (8.0, 23.4)	12.3 (4.5, 20.6)	**12.1 (4.9, 20.1)**
	4	**-282.0 (-332.8, -229.4)**	-283.7 (-333.4, -231.2)	-283.7 (-333.4, -231.2)	14.0 (8.5, 21.1)	11.4 (1.9, 19.2)	**10.9 (1.0, 18.7)**	14.9 (9.3, 23.6)	12.5 (2.0, 20.6)	**11.8 (1.2, 19.8)**
	5	**-286.3 (-342.2, -234.9)**	-286.9 (-338.3, -236.3)	-286.9 (-338.3, -236.3)	14.1 (7.4, 20.7)	11.3 (2.1, 18.4)	**10.7 (1.4, 17.6)**	14.9 (8.1, 22.3)	12.4 (4.2, 20.4)	**11.7 (2.8, 19.2)**

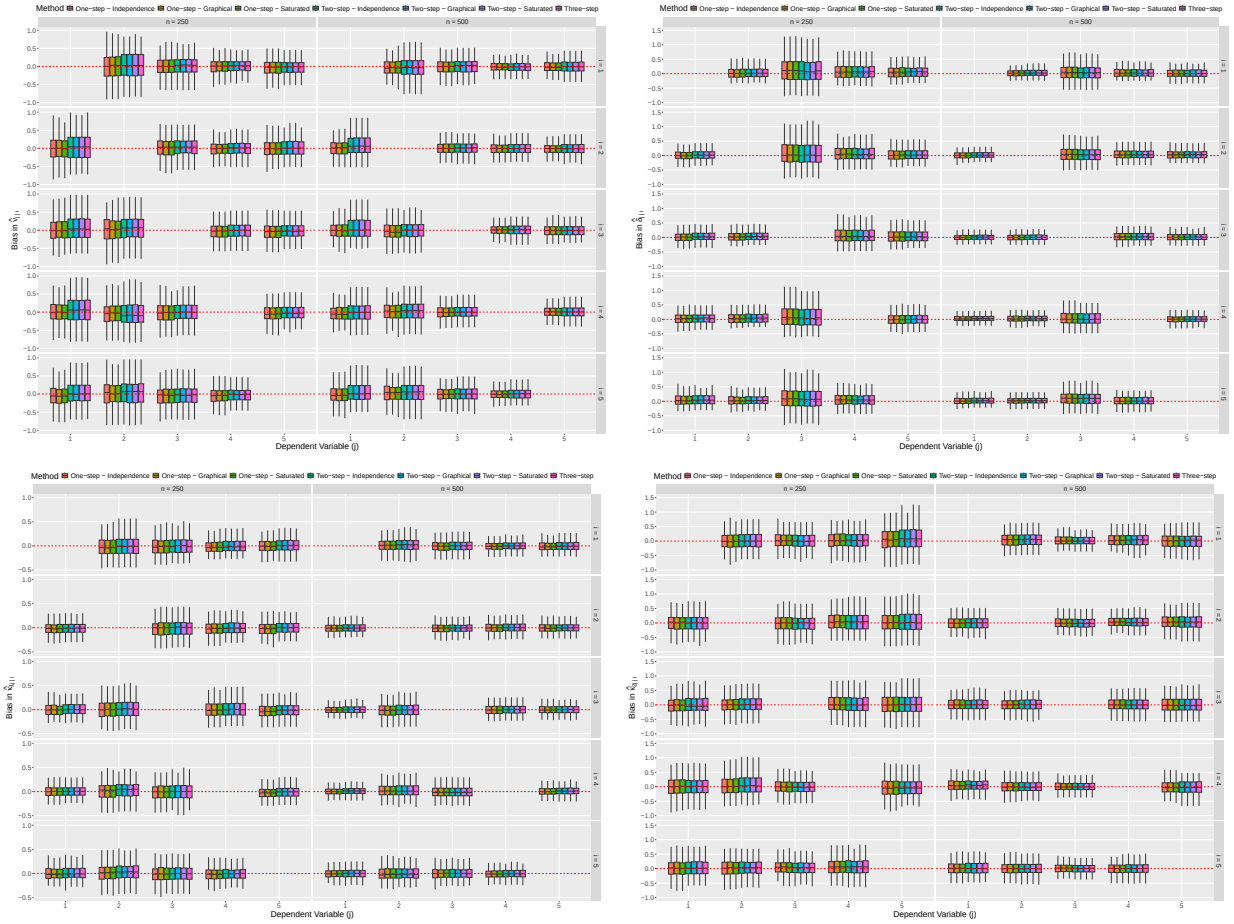

Figure S3.5: Boxplots detailing the bias of $\hat{\nu}_{j|i}$ (top left), $\hat{\delta}_{j|i}$ (top right), $\hat{\kappa}_{1_{j|i}}$ (bottom left), and $\hat{\kappa}_{2_{j|i}}$ (bottom right) for distinct $i, j \in V$. Each row corresponds to the conditioning variable i, and each column corresponds to the sample size. The different models are denoted by the fill of the boxplots. Red dashed lines show $y = 0$.

S3.3.3 Negative dependence

Similar to Section S3.3.2, we repeat the simulation from Section 4.1 of the main text but with negative associations between some components. Equation (S3.3.1) shows the true correlation matrix. All other parameters remain unchanged from Section S3.3.1.

$$\Sigma = \begin{bmatrix} 1.000 & -0.308 & -0.134 & 0.034 & 0.019 \\ -0.308 & 1.000 & -0.160 & 0.041 & 0.023 \\ -0.134 & -0.160 & 1.000 & -0.254 & -0.141 \\ 0.034 & 0.041 & -0.254 & 1.000 & -0.209 \\ 0.019 & 0.023 & -0.141 & -0.209 & 1.000 \end{bmatrix}. \tag{S3.3.1}$$

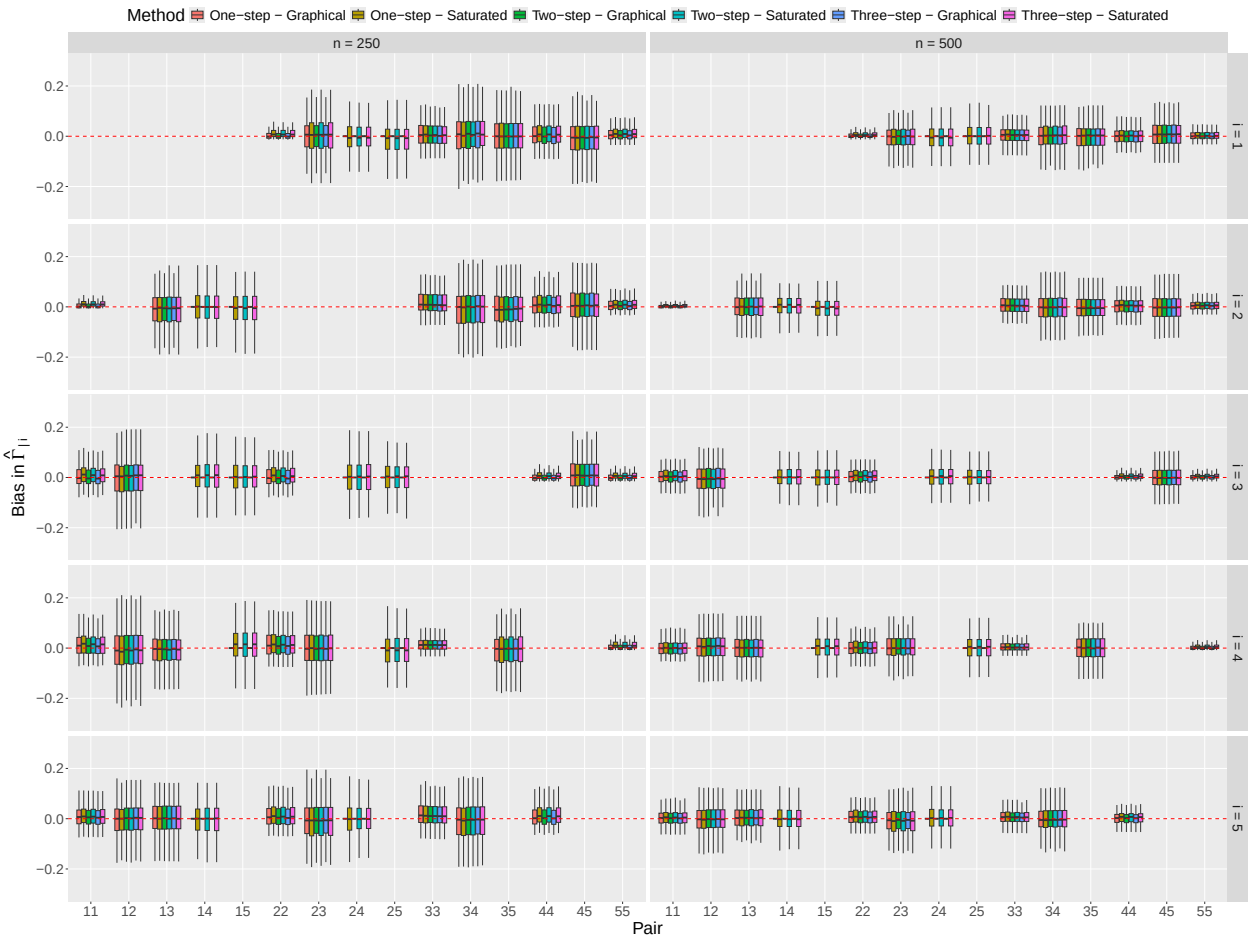

Figure S3.6: Boxplots for the bias of $\hat{\Gamma}_{|i}$ for each $i \in V$. Each row corresponds to the conditioning variable i, and each column corresponds to the sample size. The various models are denoted by the fill of the boxplots. Red dashed lines show $y = 0$.

Parameter estimates have been omitted as they are similar to those presented for the weak positive association example in the main text. To compare the stepwise inference procedures, Table S3.2 gives the biases of the fitted maximum log-likelihood values. As in the strong positive association study (Section S3.3.2), models with independent residuals have negative bias that increases with the sample size, while those with graphical or saturated dependence have small positive bias that is impervious to the sample size. Again, the magnitude of the bias is similar across all stepwise procedures, confirming no loss of information in using these.

S3.4 Majority rule proportion sensitivity analysis

In Section 3.2 of the main text, we introduce a method for selecting the optimal graphical structure. The algorithm relies on the appropriate selection of the thresholds used to fit the

Table S3.2: Median (2.5% and 97.5% quantiles) bias in the fitted maximum log-likelihood values for data from the SCMEVM with weak negative associations. Bold values denote the least biased stepwise inference procedure for each covariance structure type and conditioning variable.

Covariance Structure		Independent			Graphical			Saturated		
Number of Excesses	Conditioning Variable	One-step	Two-step	Three-step	One-step	Two-step	Three-step	One-step	Two-step	Three-step
250	1	**-13.3 (-26.7, 1.0)**	-14.2 (-27.6, -0.4)	-14.2 (-27.6, -0.4)	13.9 (8.0, 22.7)	12.5 (6.1, 21.7)	**12.4 (6.0, 21.6)**	14.9 (8.1, 27.0)	13.2 (6.8, 24.8)	**13.1 (7.0, 24.8)**
	2	**-11.6 (-28.4, 3.8)**	-13.3 (-27.6, 1.3)	-13.3 (-27.6, 1.3)	14.6 (7.5, 23.4)	**12.6 (3.3, 20.6)**	12.7 (4.8, 21.4)	15.9 (7.6, 25.1)	**13.6 (4.6, 23.5)**	**13.6 (5.5, 23.1)**
	3	**-13.3 (-46.8, 2.3)**	-14.6 (-28.7, 2.5)	-14.6 (-28.7, 2.5)	13.6 (6.9, 23.4)	11.3 (4.6, 21.0)	**11.2 (4.6, 20.8)**	15.6 (8.1, 26.3)	13.1 (5.2, 23.9)	**13.0 (5.7, 23.7)**
	4	**-12.9 (-48.7, 1.6)**	-14.3 (-26.4, -0.2)	-14.3 (-26.4, -0.2)	13.8 (7.3, 21.9)	12.3 (5.4, 20.5)	**12.2 (5.9, 20.3)**	14.6 (5.1, 23.2)	13.0 (6.7, 21.6)	**12.9 (6.8, 21.6)**
	5	**-19.7 (-35.6, -4.7)**	-21.0 (-35.7, -5.7)	-21.0 (-35.7, -5.7)	13.6 (7.3, 21.9)	**12.0 (4.7, 20.2)**	**12.0 (4.9, 20.0)**	14.7 (8.1, 23.6)	13.1 (5.7, 22.3)	**13.0 (5.6, 22.1)**
500	1	**-37.6 (-59.8, -20.8)**	-38.8 (-61.0, -22.2)	-38.8 (-61.0, -22.2)	14.4 (8.9, 22.7)	12.9 (6.1, 20.8)	**12.8 (6.0, 20.7)**	15.1 (9.4, 23.2)	13.8 (6.9, 21.5)	**13.7 (6.8, 21.4)**
	2	**-34.9 (-57.8, -12.8)**	-36.9 (-59.3, -13.1)	-36.9 (-59.3, -13.1)	14.3 (7.7, 23.1)	12.3 (4.4, 21.7)	**12.2 (4.3, 21.5)**	15.1 (8.5, 24.0)	13.2 (5.4, 22.8)	**13.1 (5.3, 22.5)**
	3	**-33.8 (-51.9, -15.4)**	-35.4 (-54.2, -17.1)	-35.4 (-54.2, -17.1)	13.1 (6.9, 20.3)	10.7 (2.1, 18.6)	**10.6 (2.5, 18.5)**	14.8 (8.2, 23.0)	**12.3 (3.1, 20.3)**	**12.3 (3.0, 20.2)**
	4	**-40.5 (-57.8, -19.5)**	-41.7 (-59.7, -20.9)	-41.7 (-59.7, -20.9)	13.9 (8.2, 22.9)	11.9 (5.1, 20.6)	**11.7 (5.0, 20.5)**	14.7 (9.1, 24.1)	12.8 (5.4, 21.9)	**12.7 (5.4, 21.7)**
	5	**-51.2 (-70.9, -31.4)**	-52.8 (-70.6, -33.2)	-52.8 (-70.6, -33.2)	13.8 (8.5, 22.6)	12.0 (5.3, 21.3)	**11.9 (5.1, 20.9)**	14.5 (9.3, 23.9)	12.9 (6.4, 22.3)	**12.7 (6.2, 22.0)**

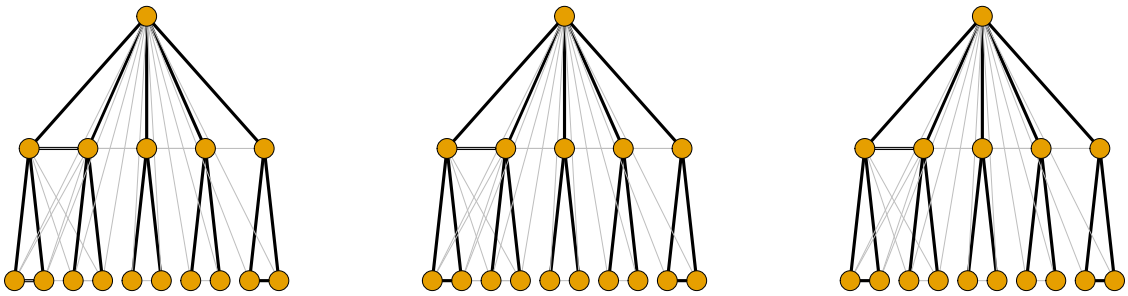

Figure S3.7: Inferred graphical structure for data generated from the multivariate Pareto distribution when the majority rule proportion p in Algorithm 3.5 of the main text is set to 0.3 (left), 0.5 (centre), and 0.7 (right). The line width and darkness in each panel correspond to the number of times each edge was selected across 100 samples. Black and grey edges correspond to true and additional edges, respectively.

CMEVMs and the majority-rule proportion used to obtain the final unified structure (step 11 of Algorithm 3.5). In Section 4.2 of the main text, we explored sensitivity to threshold choice. We now assess sensitivity to the majority-rule proportion.

Continuing the simulation study in Section 4.2 of the main text, we rerun Algorithm 3.5 with u_{Y_i} set to the 0.8-quantile of the standard Laplace distribution for all $i \in V$. We then compare results using three majority-rule proportions, $p \in \{0.3, 0.5, 0.7\}$. Figure S3.7 shows weighted graphs, with line width and darkness proportional to the number of times the edge is selected across the 100 datasets, for $p = 0.3$ (left), $p = 0.5$ (centre) and $p = 0.7$ (right). The true graph is identified in all cases with very few incorrect (grey) edges. This suggests that the graphical selection algorithm is not overly sensitive to the choice of majority-rule proportion. Nevertheless, it is important to include it to remove spurious edges that may only occur in one of the conditioned subgraphs.

To check that these results are not confined to the multivariate Pareto distribution, we repeat

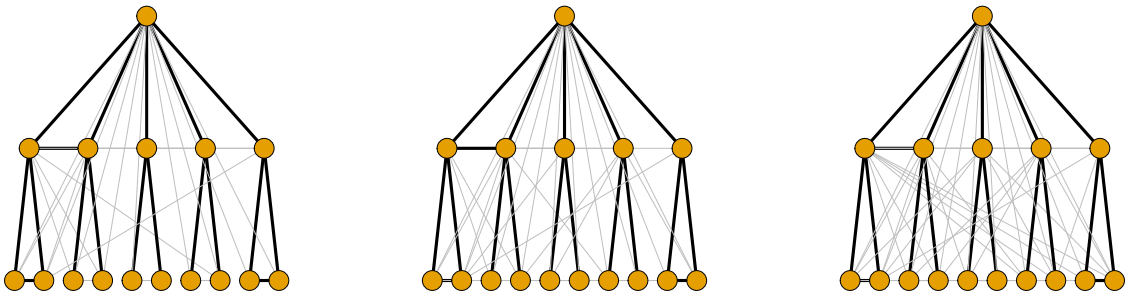

Figure S3.8: Inferred graphical structure for data generated from the multivariate Gaussian distribution when the majority rule proportion p in Algorithm 3.5 of the main text is set to 0.3 (left), 0.5 (centre), and 0.7 (right). The line width and darkness in each panel correspond to the number of times each edge was selected across 100 samples. Black and grey edges correspond to true and additional edges, respectively.

the simulation study, taking the underlying generating mechanism to be the multivariate Gaussian distribution. Taking $d = 16$ and $\mathcal{G}$ as in the left panel of Figure 3 in the main text, we simulate $1,000$ points from a multivariate Gaussian distribution with the components of the mean vector $\boldsymbol{\mu}$ independently sampled from a uniform distribution on $(-5, 5)$ and correlation matrix consistent with $\mathcal{G}$. We use 100 replicates for this study.

To infer the optimal graphical structure, we use Algorithm 3.5 of the main text. In the algorithm, we set the dependence thresholds u_{Y_i} to the 0.90-quantile of the standard Laplace distribution for all $i \in V$. Figure S3.8 shows weighted graphs of the 100 inferred graphical structures, with line width and darkness proportional to the number of times the edge is selected across the 100 datasets, when the majority rule proportion is set to 0.3 (left), 0.5 (centre), and 0.7 (right). Again, the true graph is well identified with very few "incorrect" (grey) edges.

Table S3.3 gives frequency counts of the number of edges inferred by Algorithm 3.5 of the main text, for both the multivariate Pareto and the MVG generating mechanisms. For the multivariate Pareto distribution, the distribution of edges in the selected graphs is very similar across all three proportions, suggesting the more important tuning parameter is the threshold above which the CMEVMs are fitted. For the MVG generating mechanism, increasing the majority rule proportion shifts the distribution towards the true number of edges. This is expected since edges are included in the final graph only if they are important most of the time.

Table S3.3: Number of times, out of the 100 samples, a graph with x edges is inferred using Algorithm 3.5 of the main text for various majority rule proportions and underlying generating mechanisms.

Generating Mechanism	Majority Rule Proportion	Number of edges in $\mathcal{G}$											
		17	18	19	20	21	22	23	24	25	26	27	28
MVP	0.3	6	33	24	16	15	4	2	-	-	-	-	-
	0.5	7	21	27	24	9	9	-	1	2	-	-	-
	0.7	4	24	25	23	14	6	2	2	-	-	-	-
MVN	0.3	2	30	22	20	10	5	5	3	1	1	1	-
	0.5	-	38	29	14	8	7	2	2	-	-	-	-
	0.7	-	42	29	15	5	3	1	3	-	1	-	1

S3.5 Additional graph selection example

In Section 4.2 of the main text, we replicate the simulation study of Engelke and Hitz (2020, Section 5.5) to assess how closely the SCMEVM can identify the graphical structure of data generated from the Hüsler-Reiss distribution. Here, we repeat the study for data generated from the multivariate Gaussian distribution; see Section S3.4 for details of data simulation.

For the SCMEVM, we use Algorithm 3.5 of the main text to infer the optimal graphical structure, setting the dependence thresholds u_{Y_i} to the 0.90-quantile of the standard Laplace distribution for all $i \in V$ and the majority rule proportion to 0.5. Figure S3.9 (centre panel) shows a weighted graph of inferred graphical structures with line width and darkness proportional to the number of times the edge is selected across 100 simulated datasets. The true graphical structure is clearly recovered.

For comparison, we also use the graphical selection method EGlearn (Engelke et al., 2025). EGlearn uses a multivariate Pareto model but generalises the allowable graphical structure beyond the block structure requirement in Engelke and Hitz (2020). The only tuning parameter is the threshold above which the data are assumed to follow the multivariate Pareto distribution. We set this to be the 0.95-quantile of the standard Pareto distribution. The resulting weighted graph over the 100 datasets is shown in Figure S3.9 (right panel). The true graphical structure is largely identified, but there are more "incorrect" edges compared to the SCMEVM method. This result seems counter-intuitive since EGlearn is designed to learn the graphical structure for AD data and not AI data. That being said, the results are somewhat threshold sensitive. Using a lower threshold of the 0.8-quantile of the standard

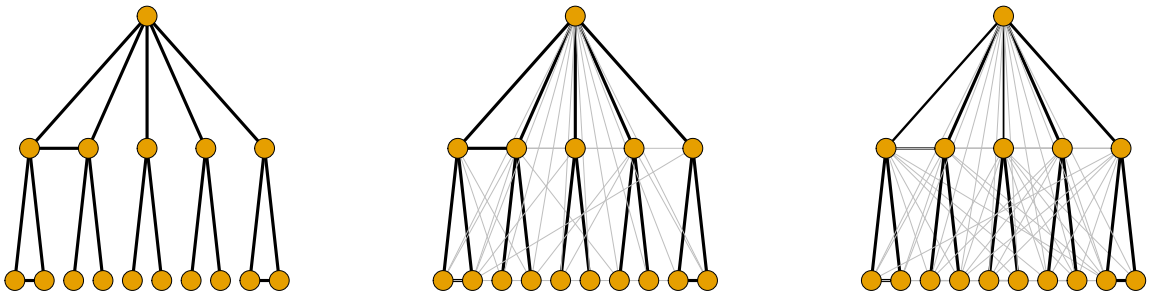

Figure S3.9: True underlying graphical structure (left) and inferred graphical structures using the method proposed in Section 3.2 of the main text (centre) and using "EGlearn" (right). Line width and darkness indicate the number of times each edge was selected across 100 replicates. Black and grey edges correspond to "true" and "additional" edges, respectively.

Pareto distribution results in 5 additional edges in the pruned weighted graph, while a higher threshold of the 0.99-quantile of the standard Pareto distribution results in six missing edges in the pruned weighted graph and the algorithm inferring fewer than 18 edges in 97% of cases. In contrast, our method is relatively stable with respect to the threshold, as it identifies the true graph when u_{Y_i} is set to the $0.8-$ and $0.95-$quantile of the standard Laplace distribution for all $i \in V$.

These results suggest that the Engelke et al. (2025) method can struggle to correctly identify the underlying structure of the data when the data-generating mechanism does not have complete AD, while the method proposed in the main text does not have this limitation.

S3.6 Additional figures and simulation studies for Section 4.3

In the main text, we considered data with a mixture of extremal dependence structures. We now repeat the simulation study in Section 4.3 of the main text for data that exhibits either full asymptotic independence (AI) or full asymptotic dependence (AD). For AI, we simulate from each of the (a) multivariate Gaussian (MVG), (b) symmetric multivariate Laplace (MVL), and (c) multivariate t- (MVT) distributions. In all cases, both positive and negative associations are investigated. For AD, we simulate from a multivariate Pareto (MVP) distribution.

All simulation studies follow a similar pattern. For each true distribution, 200 datasets are sampled using a dependence structure consistent with $\mathcal{G}$ in Section S3.3. Data are trans-

formed from their canonical margins (e.g., Gaussian for the MVG distribution) to standard Laplace margins as per Section 3.1 of the main text. For all datasets, each of the one-, two- and three-step procedures is used to fit the SCMEVM with graphical covariance, where the graph is assumed to be known and correctly specified. The three-step procedure is also used to fit the SCMEVM with independent and saturated covariances. For comparison, we also fit the original CMEVM (Heffernan and Tawn, 2004) and the graphical extremes model (Engelke and Hitz, 2020) (EHM). The mean absolute error (MAE) and root mean squared error (RMSE) of the model-based estimates form the basis of model comparison. Such metrics are calculated via probabilities of the form $\mathbb{P}[\boldsymbol{X}_A > u_{\boldsymbol{X}_A} \mid X_i > u_{X_i}]$ ($\mathbb{P}[\boldsymbol{X}_A < u_{\boldsymbol{X}_A} \mid X_i > u_{X_i}]$) for simulations that have positive (negative) associations and for all sets $A \subseteq V_{|i}$ and $i \in V$

S3.6.1 Multivariate Gaussian distribution

In this section, we assume $\boldsymbol{X}$ follows a MVG distribution with mean vector $\boldsymbol{\mu}$, where each μ_j is independently sampled from a uniform distribution on $(-5, 5)$, and correlation matrix Σ. Various strengths of correlation are considered, however $\Gamma = \Sigma^{-1}$ is always consistent with $\mathcal{G}$ in Section S3.3. We set dependence thresholds u_{Y_i} to the 0.90-quantile of the standard Laplace distribution for all $i \in V$. For prediction, we set u_{X_i} to the 0.95-quantile for the true distribution of X_i for each $i \in V$.

Weak positive dependence

In the first study, correlations between all components lie in $(0, 0.47)$. Figure S3.10 shows MLEs of the dependence and AGG parameters. Here, and in the other studies in this section, estimates from the three-step SCMEVM with graphical and saturated covariance structures are omitted as they are identical to results for the three-step SCMEVM with independent residuals. Also note that the MLEs for CMEVM dependence parameters are the same for the two- and three-step methods. The MLEs of $\boldsymbol{\alpha}_{|i}$ ($\boldsymbol{\nu}_{|i}$) from the one-step procedure are consistently lower (higher) than the MLEs from the stepwise approaches, confirming that the one-step method does not guarantee that the first-order extremal dependence structure will be captured by the dependence parameters. At best, by attributing some extremal dependence structure to the residual distribution, the interpretability of the SCMEVMs fitted with the one-step procedure is reduced. Potentially, it also makes the models less reliable. In contrast, the MLEs of $\boldsymbol{\beta}_{|i}$, $\boldsymbol{\kappa_1}_{|i}$, $\boldsymbol{\kappa_2}_{|i}$, and $\boldsymbol{\delta}_{|i}$ are similar across all the models and fitting procedures.

Note that the estimates for $\beta_{j|i}$ should be close to 0.5 when the underlying generating mech-

anism is the MVG, and there is a positive association between components i and j. From the top right panel of Figure S3.10, this is clearly not the case. Since the estimated parameters for the one- and two-/three-step fits are similar, the underestimation is likely due to the slow convergence of the MVG distribution to its asymptotic limit. To improve convergence, one would likely need to simulate a larger dataset (currently set at $5,000$) and/or use a lower dependence threshold (currently set to the $0.9-$ quantile of the standard Laplace distribution). Another thing to note is the right-scale parameter $\kappa_{2_{j|i}}$ is almost always estimated to be larger than the left-scale parameter $\kappa_{1_{j|i}}$ for distinct $i, j \in V$, supporting the choice of an asymmetric marginal distribution for $\boldsymbol{Z}_{|i}$.

Figure S3.11 shows empirical and model-based estimates of the conditional precision matrix $\Gamma_{|i}$. Empirical estimates are the inverse of the conditional correlation matrix for $\boldsymbol{Y} \mid Y_i = y_i$, such that $y_i > u_{Y_i}$, equivalently the inverse correlation matrix of $\boldsymbol{Y} \mid Y_i > u_{Y_i}$ excluding the ith row and column. Similar to the study in the main text, the estimated matrices are the same for the graphical and saturated SCMEVMs, confirming there is negligible loss in using the former. Further, the estimated structure of the conditional precision matrices for the graphical and saturated SCMEVMs is consistent with the empirical version. While it is plausible that the results here are specific to the MVG generating mechanism, similar patterns are observed for the other multivariate distributions.

We now compare predictions from the EHM and three-step SCMEVM with graphical covariance. Figure S3.12 (left panel) shows the bias in the conditional survival curves of $X_j \mid X_1 > u_{X_1}$ for each $j \in V_{|1}$. The SCMVEM is unbiased for all curves, whereas the EHM has positive bias for lower values of u_{X_j}; this decreases as u_{X_j} increases. The positive bias of the EHM persists in bivariate conditional survival probabilities. Figure S3.12 (right panel) shows the bias in $\mathbb{P}[X_2 > u_{X_2}, X_3 > u_{X_3} \mid X_1 > u_{X_1}]$. The three-step SCMEVM with independent residuals exhibits negative bias because X_2 is not conditionally independent of X_3 given X_1. In contrast, the SCMEVMs with graphical and saturated covariances are unbiased. The CMEVM predictions are also unbiased due to the low dimension d. Lastly, the SCMEVMs with graphical covariance exhibit the least amount of bias and variability, minimising the MAE and RMSE for 87% and 77% of the 75 conditional probabilities, respectively. This confirms that there is no loss in performance when using a graphical structure over the more flexible saturated one and that the fully parametric SCMEVM outperforms the semi-parametric CMEVM. The EHM performs poorly in this case because the true data have AI.

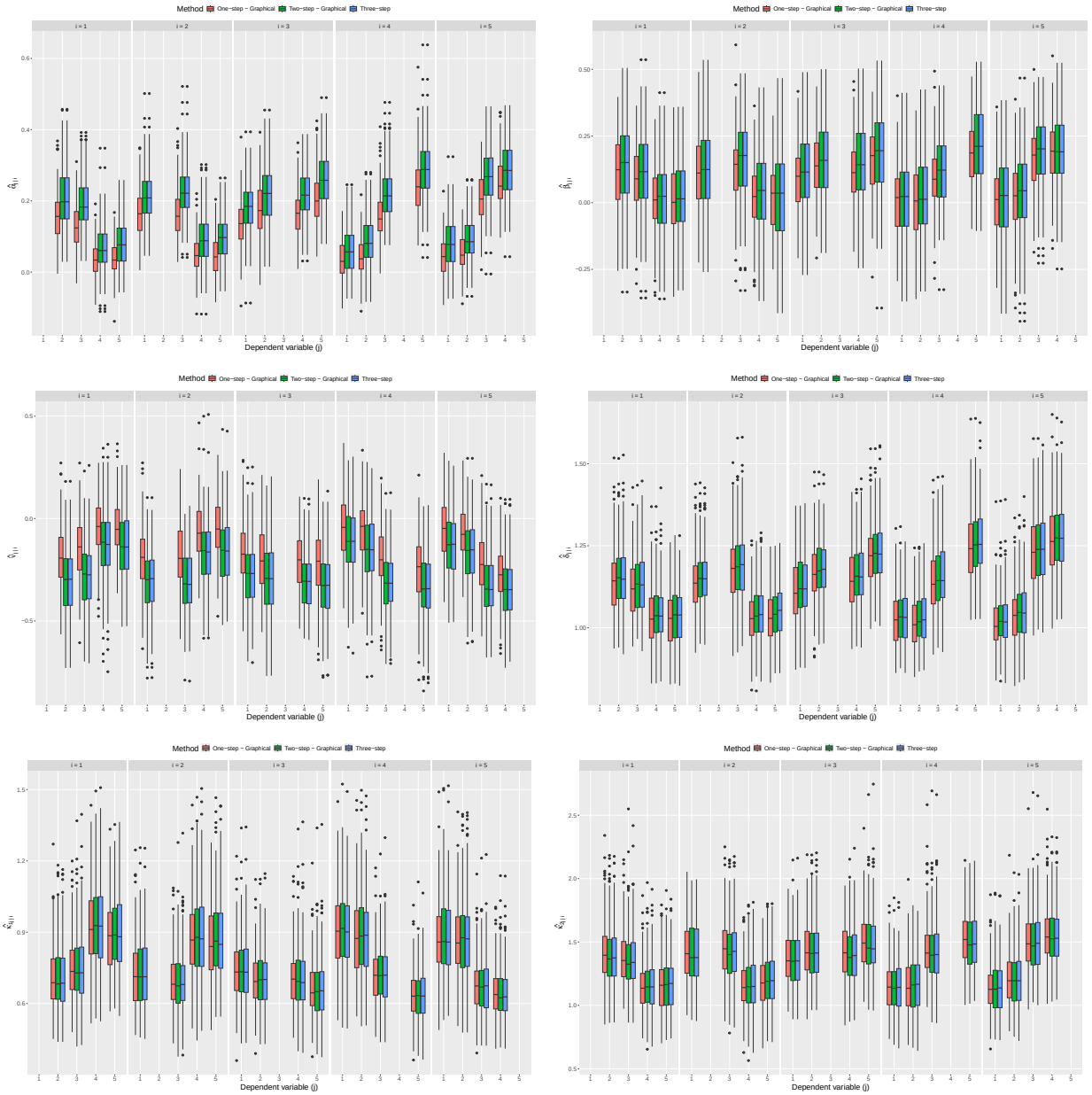

Figure S3.10: Boxplots of MLEs for $\alpha_{j|i}$ (top left), $\beta_{j|i}$ (top right), $\nu_{j|i}$ (centre left), $\delta_{j|i}$ (centre right), $\kappa_{1_{j|i}}$ (bottom left), and $\kappa_{2_{j|i}}$ (bottom right) for distinct $i, j \in V$. Each column corresponds to the conditioning variable i. The different models are denoted by the fill of the boxplots.

Strong positive dependence

We repeat the simulation study in Section S3.6.1, but the associations between the components of $\boldsymbol{X}$ are strong and positive (> 0.52). We present only the predictive performances, as the parameter estimates show similar patterns to those seen in Section S3.6.1. Figure S3.13 (left panel) shows bias in the conditional survivor curves for $X_j \mid X_5 > u_{X_5}$ such that $j \in V_{|5}$

Figure S3.11: Boxplots of empirical and model-based estimates of $\Gamma_{|i}$, for each $i \in V$, when the data is generated from a MVG distribution with weak positive associations. Each row corresponds to the conditioning variable i, and each column corresponds to the correlation parameter. The different models are denoted by the colour of the boxplots. Black dashed lines show $y = 0$.

from both the EHM and the three-step SCMEVM with a graphical covariance structure. Again, the EHM is biased for low values of u_{X_j}, but this diminishes as u_{X_j} increases; the three-step SCMEVM with graphical structure is unbiased for all u_{X_j}. Figure S3.13 (right panel) shows the bias in $\mathbb{P}[\boldsymbol{X}_{|5} > u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$. The EHM has positive bias, whereas both the CMEVM and the SCMEVMs with graphical or saturated covariance structures are unbiased. The three-step SCMEVM with independent residuals exhibits negative bias; this is expected since the components of $\boldsymbol{X}_{|5}$ are not independent given X_5 is large. Assessing overall predictive performance, the SCMEVMs with graphical covariance structure are again the least biased and variable, minimising the MAE and RMSE metrics 81% and 84% of the time, respectively.

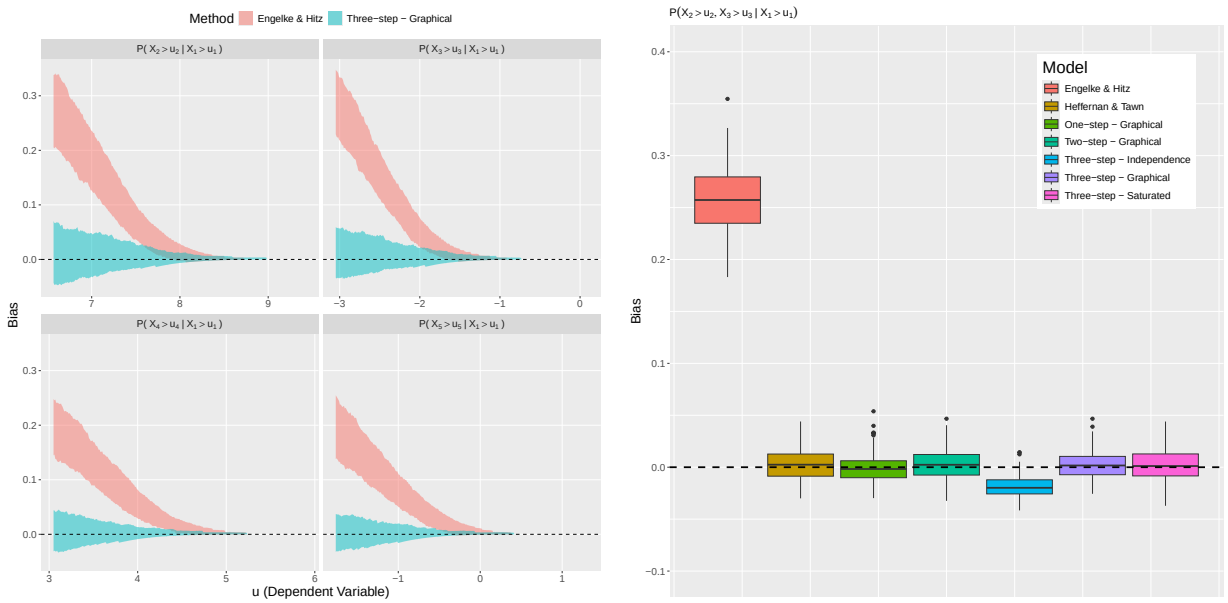

Figure S3.12: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j > u_{X_j} \mid X_1 > u_{X_1}]$, for each $j \in V_{|1}$, where $\boldsymbol{X}$ follows a MVG distribution with weak positive associations (left). The bias from the EHM and the three-step SCMEVM, assuming a graphical covariance structure for the residuals, are in pink and blue, respectively. Boxplots of the bias in $\mathbb{P}[X_2 > u_{X_2}, X_3 > u_{X_3} \mid X_1 > u_{X_1}]$ (right). The bias from the various models is denoted by the fill of the boxplots. Black dashed lines show $y = 0$.

Negative dependence

We repeat the simulation study in Section S3.6.1, but the association between the components of $\boldsymbol{X}$ are now allowed to be negative. The correlation matrix Σ is given in equation (S3.6.1).

$$\Sigma = \begin{bmatrix} 1.000 & -0.468 & -0.370 & -0.136 & 0.134 \\ -0.468 & 1.000 & 0.390 & 0.144 & -0.141 \\ -0.370 & 0.390 & 1.000 & 0.369 & -0.362 \\ -0.136 & 0.144 & 0.369 & 1.000 & -0.346 \\ 0.134 & -0.141 & -0.362 & -0.346 & 1.000 \end{bmatrix}. \tag{S3.6.1}$$

Figure S3.14 compares the MLEs of $\kappa_{1_{j|i}}$ and $\kappa_{2_{j|i}}$ from the three-step SCMEVM with a graphical covariance structure, for distinct $i, j \in V$. In the other MVG examples, the right-scale parameter is generally larger than the left-scale parameter, whereas here a range of behaviour is observed. This further justifies the need for a flexible, asymmetric distribution for $\boldsymbol{Z}_{|i}$.

Figure S3.15 (left panel) shows the bias in the conditional cumulative distribution curves for $X_j \mid X_5 > u_{X_5}$ for the EHM and the three-step SCMEVM with graphical covariance structure.

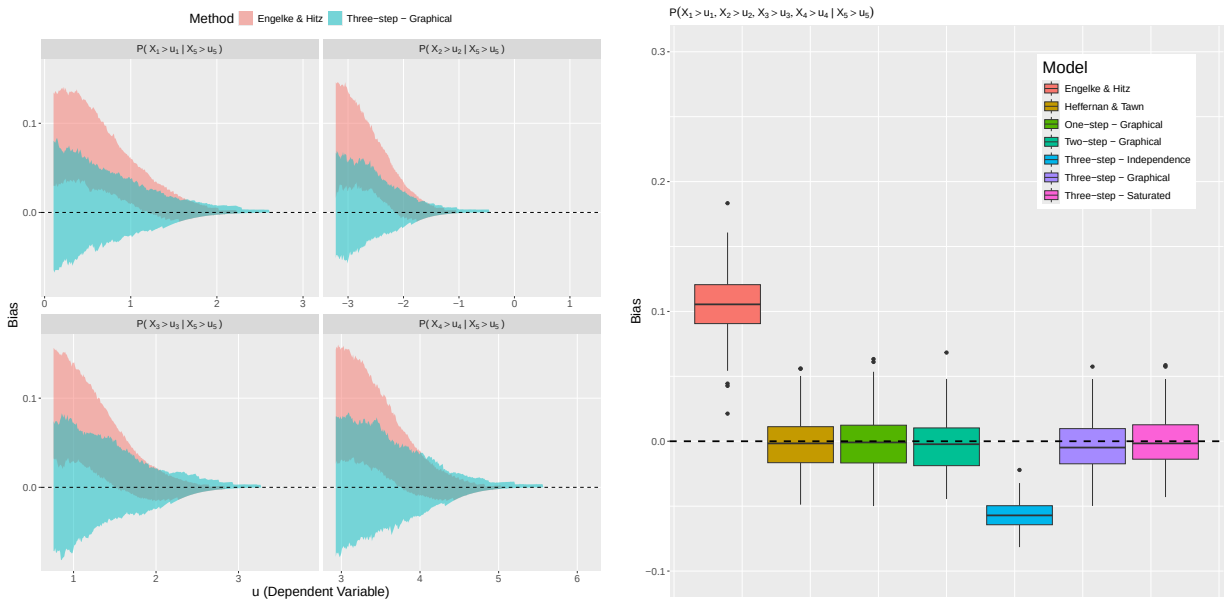

Figure S3.13: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j > u_{X_j} \mid X_5 > u_{X_5}]$, for $j \in V_{|5}$ where $\boldsymbol{X}$ follows a MVG distribution with strong positive associations (left). The bias from the EHM and the three-step SCMEVM with a graphical covariance structure are in pink and blue, respectively. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|5} > u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$ (right). The bias from the various models is denoted by the fill of the boxplots. Black dashed lines show $y = 0$.

The three-step SCMEVM exhibits no bias, but the EHM underestimates the curve over the entire range. Again, this is not surprising, as the AD assumption is not satisfied by the data. Figure S3.15 (right panel) considers the bias in $\mathbb{P}[\boldsymbol{X}_{|5} < u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$. The EHM exhibits negative bias, while the CMEVM and SCMEVMs are unbiased. Finally, as in the previous studies, the SCMEVMs with graphical covariance structures are the least biased and variable, minimising the MAE and RMSE 76% and 87% of the time, respectively.

S3.6.2 Multivariate Laplace distribution

In this study, $\boldsymbol{X}$ follows a MVL distribution with mean vector $\boldsymbol{\mu}$, where μ_j are independently sampled from a uniform distribution on $(-5, 5)$, and precision matrix consistent with $\mathcal{G}$ in Section S3.3.

Weak positive dependence

In this simulation, associations between components are weakly positive i.e., the elements of the true correlation matrix are strictly positive but less than 0.37. We set the dependence threshold u_{Y_i} to the 0.95-quantile of the standard Laplace distribution for all $i \in V$, resulting

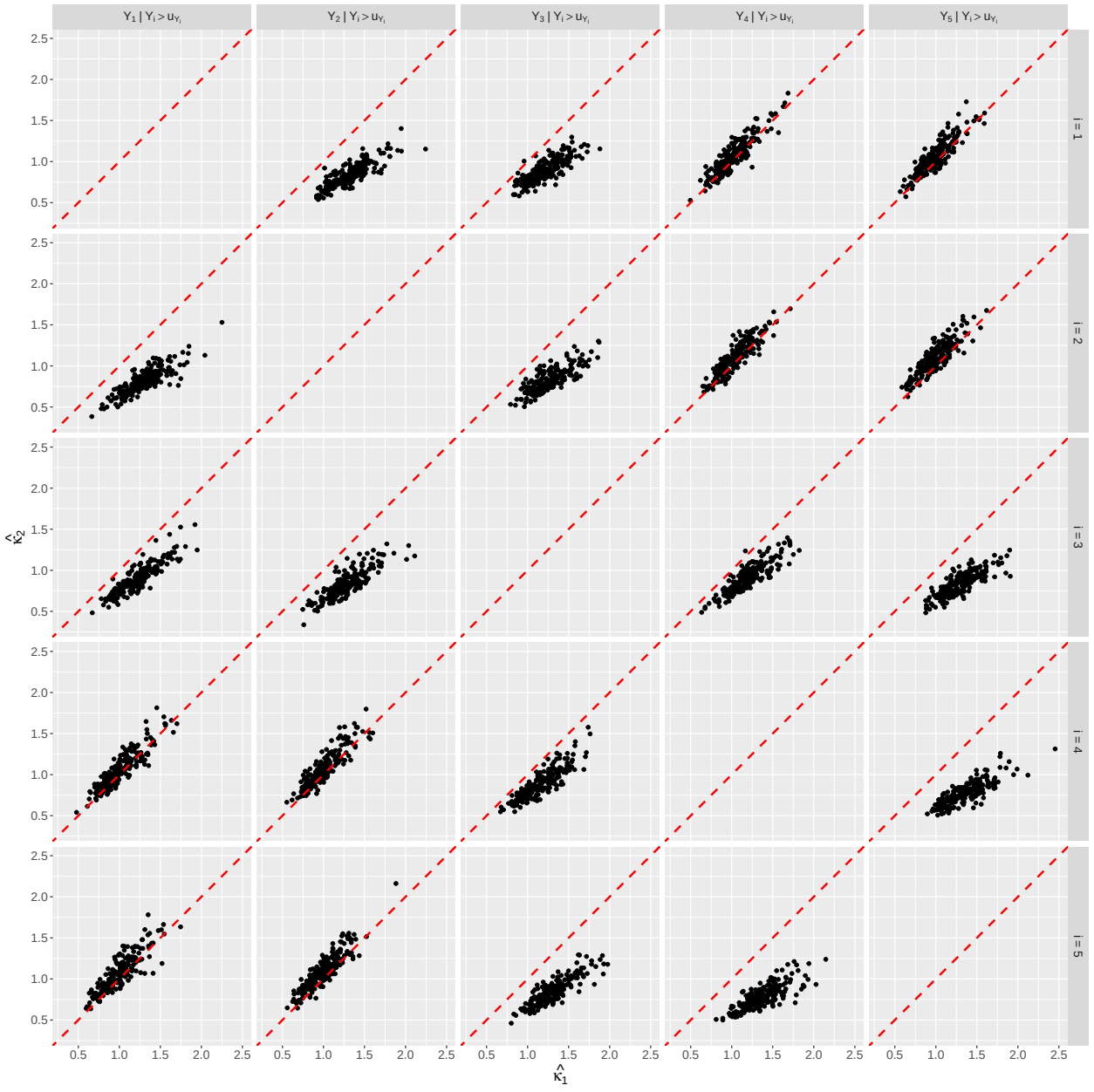

Figure S3.14: Scatter plots comparing $\hat{\kappa}_{1_{j|i}}$ and $\hat{\kappa}_{2_{j|i}}$ from the three-step SCMEVM with graphical covariance structure for distinct $i, j \in V$. Red dashed lines show $y = x$.

in approximately 250 excesses per conditioning variable. For prediction, u_{X_i} is set to the 0.95-quantile from a single sample of 10^6 from the true distribution. Figure S3.16 shows empirical and model-based estimates of the conditional precision matrix $\Gamma_{|i}$. The estimated structure of the conditional precision matrix from both the graphical and the saturated SCMEVM is consistent with the empirical version. Analysis of other parameter estimates has been omitted, but they are very similar across all three stepwise procedures. The only point to note is that estimates for the left- and right-scale parameters in the MVAGG are very similar,

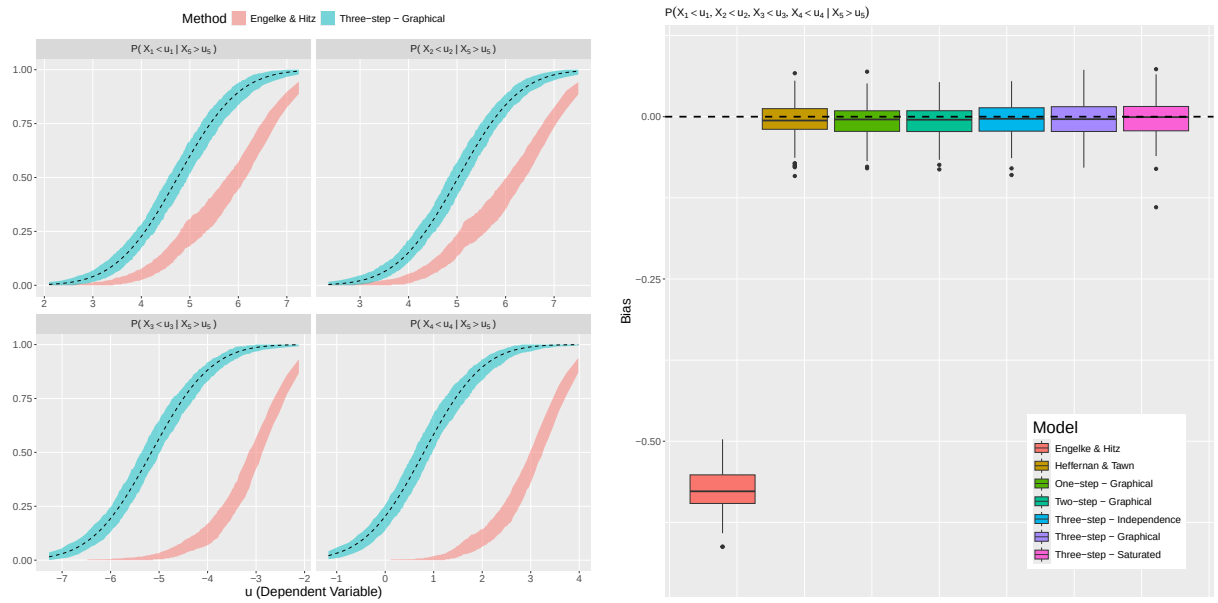

Figure S3.15: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of $\mathbb{P}[X_j < u_{X_j} \mid X_5 > u_{X_5}]$ for $j \in V_{|5}$, when $\boldsymbol{X}$ follows a MVG distribution with negative associations (left). The estimated curves from the EHM and the three-step SCMEVM with a graphical covariance structure are in pink and blue, respectively. The true conditional cumulative distribution curves are given by the black dashed lines. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|5} < u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$ (right). The bias from the various models is denoted by the fill of the boxplots. The $y = 0$ line is indicated by the black dashed line.

raising the question of whether the generalised Gaussian would be a better choice of marginal residual distribution. However, other examples do have very different scale parameters (see Figures S3.10 and S3.14) and the more flexible asymmetric generalised Gaussian distribution is therefore recommended.

To compare predictive performance, Figure S3.17 (left panel) shows the bias in the conditional survival curves of $X_j \mid X_3 > u_{X_3}$ for $j \in V_{|3}$. Similar to the MVG examples, the SCMEVM with graphical covariance structure is unbiased for all curves, whereas the EHM has positive bias for low values of u_{X_j}, which decreases as u_{X_j} increases. Figure S3.17 (right panel) shows the bias in $\mathbb{P}[\boldsymbol{X}_{|3} > u_{\boldsymbol{X}|3} \mid X_3 > u_{X_3}]$. The CMEVM and the SCMEVMs with graphical or saturated covariance structures are unbiased, whereas the EHM has positive bias. Again, the three-step SCMEVM with independent residuals has negative bias because not all components of $\boldsymbol{X}_{|3}$ are independent given X_3. Similar findings are made when assessing other conditional probabilities of the form $\mathbb{P}[\boldsymbol{X}_A > u_A \mid X_i > u_i]$ for all $A \subseteq V_{|i}$ and $i \in V$. Lastly, the SCMEVMs with graphical covariance have the least amount of bias and variability, minimising the MAE and RMSE for 86% and 77% of the 75 conditional probabilities, respectively.

Figure S3.16: Boxplots of empirical and model-based estimates of $\Gamma_{|i}$, for each $i \in V$, when the data is generated from a MVL distribution with weak positive associations. Each row corresponds to the conditioning variable i, and each column corresponds to the correlation parameter. The different models are distinguished by the colour of the boxplots. Black dashed lines show $y = 0$.

Strong positive dependence

We repeat the simulation study in Section S3.6.2 with strong positive association between the components, i.e., the entries of the true correlation matrix are all greater than 0.69. The dependence threshold u_{Y_i} is set to the 0.9-quantile for the standard Laplace distribution, for all $i \in V$. For prediction, we set u_{X_i} to the 0.95-quantile from a dataset of size 10^6 simulated from the true distribution for each $i \in V$. We omit parameter estimates since the only point to note is that a comparison of the estimates from the one-, two-, and three-step SCMEVMs shows that the dependence parameters are slightly larger for the one-step method, while the location and scale parameters are slightly lower.

Figure S3.18 shows the bias in two tail probabilities, $p_1 = \mathbb{P}[X_2 > u_{X_2}, X_3 > u_{X_3} \mid X_1 > u_{X_1}]$ and $p_2 = \mathbb{P}[\boldsymbol{X}_{|1} > u_{\boldsymbol{X}_{|1}} \mid X_1 > u_{X_1}]$. In this case, the EHM performs more similarly to the CMEVM and SCMEVMs. However, the model exhibits a small positive bias for p_2.

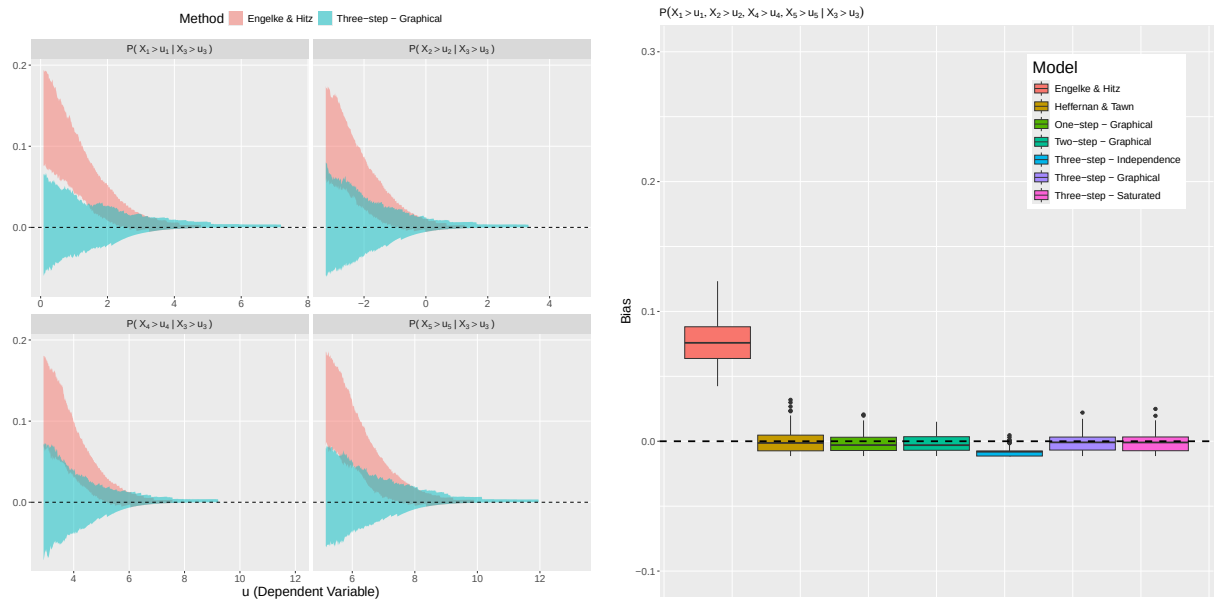

Figure S3.17: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j > u_{X_j} \mid X_3 > u_{X_3}]$ for $j \in V_{|3}$, where $\boldsymbol{X}$ follows a MVL distribution with weak positive associations (left). The bias from the EHM and the three-step SCMEVM, assuming a graphical covariance structure for the residuals are in pink and blue, respectively. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|3} > u_{\boldsymbol{X}_{|3}} \mid X_3 > u_{X_3}]$ (right). The bias from the various models is denoted by the fill of the boxplots. Black dashed lines show $y = 0$.

The EHM minimises the MAE and RMSE for 37 and 42 of the 75 conditional probabilities, respectively. For comparison, the three-step SCMEVM with graphical covariance structure minimises the metrics 31 and 25 times, respectively. Despite poorer performance on the metrics, the CMEVM and SCMEVMs with graphical or saturated covariance structures are unbiased for both probabilities in Figure S3.18, suggesting these models scale better with dimension compared to the EHM.

Negative dependence

Finally, we consider negative associations between components. The true correlation matrix is provided in equation (S3.6.2). The dependence threshold u_{Y_i} is set to the 0.8-quantile for the standard Laplace distribution for all $i \in V$. For prediction, we set u_{X_i} to the 0.9-quantile from a dataset of size 10^6 simulated from the true distribution for each $i \in V$.

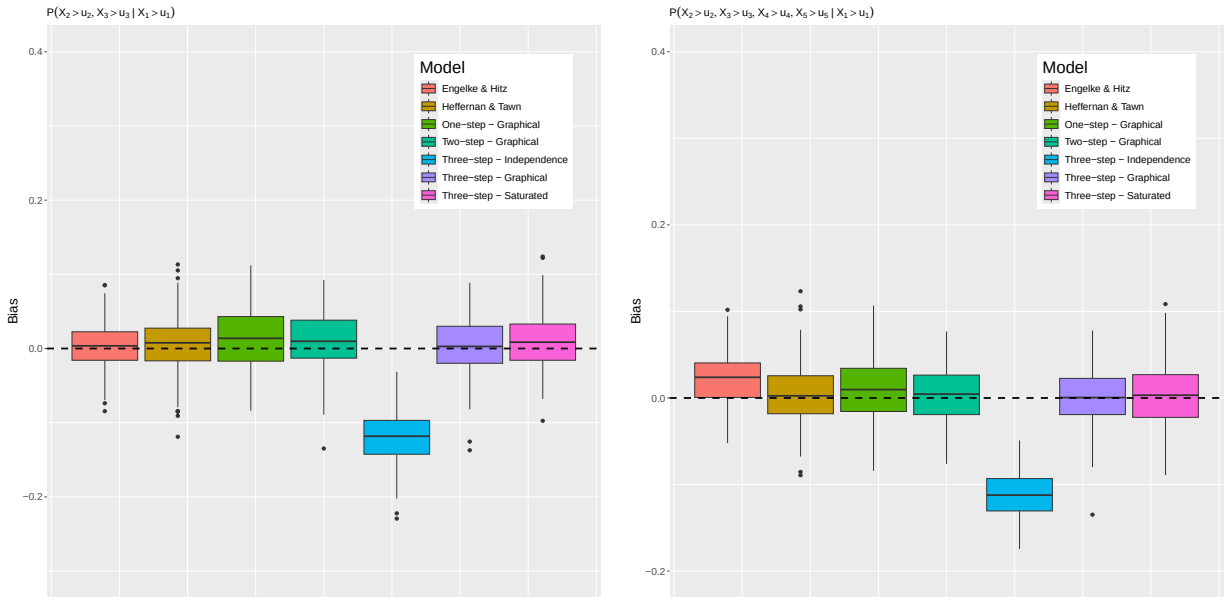

Figure S3.18: Boxplots of the bias in $p_1 = \mathbb{P}[X_2 > u_{X_2}, X_3 > u_{X_3} \mid X_1 > u_{X_1}]$ (left) and $p_2 = \mathbb{P}[\boldsymbol{X}_{|1} > u_{\boldsymbol{X}_{|1}} \mid X_1 > u_{X_1}]$ (right) when $\boldsymbol{X}$ follows a MVL distribution with strong positive associations. The bias from the various models is denoted by the fill of the boxplots. Black dashed lines show $y = 0$.

$$\Sigma = \begin{bmatrix} 1.000 & -0.200 & -0.139 & 0.026 & 0.022 \\ -0.200 & 1.000 & -0.243 & 0.045 & 0.038 \\ -0.139 & -0.243 & 1.000 & -0.185 & -0.158 \\ 0.026 & 0.045 & -0.185 & 1.000 & -0.276 \\ 0.022 & 0.038 & -0.158 & -0.276 & 1.000 \end{bmatrix}. \tag{S3.6.2}$$

The only point to note on the parameter estimates is that $\beta_{j|i}$ tends to always be slightly higher for the one-step SCMEVM than for the two- and three-step SCMEVMs. Figure S3.19 (left panel) shows 95% confidence intervals for the conditional cumulative distribution curves of $X_j \mid X_4 > u_{X_4}$ from the EHM and the three-step SCMEVM with graphical covariance structure for $j \in V_{|4}$. As with the MVG distribution with negative association, the SCMEVM captures the curves perfectly, while the EHM underestimates all curves. The SCMEVMs with graphical covariance structure minimise the MAE and RMSE 63% and 59% of the time, respectively. The three-step SCMEVM with saturated covariance structure also performs very well, minimising the metrics 23% and 28% of the time, respectively. However, the numerical values of the metrics are almost indistinguishable for the two models as shown in Figure S3.19 (right panel) where we plot the bias in $\mathbb{P}[\boldsymbol{X}_{|4} < u_{\boldsymbol{X}_{|4}} \mid X_4 > u_{X_4}]$. The EHM and SCMEVM with independent residuals are both biased, while the SCMEVMs with

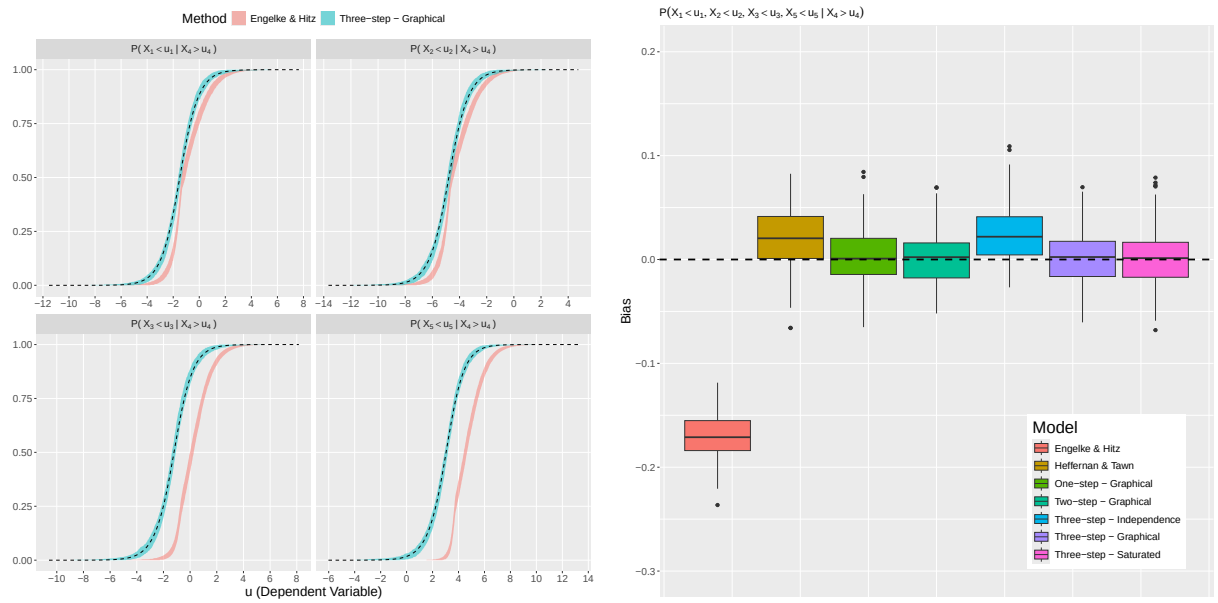

Figure S3.19: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j < u_{X_j} \mid X_4 > u_{X_4}]$ for $j \in V_{|4}$. where $\boldsymbol{X}$ follows a MVL distribution with negative associations (left). The bias from the EHM and the three-step SCMEVM, assuming a graphical covariance structure for the residuals are in pink and blue, respectively. The true conditional cumulative distribution curves are given by the black dashed lines. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|4} < u_{\boldsymbol{X}_{|4}} \mid X_4 > u_{X_4}]$ (right). The bias from the various models is denoted by the fill of the boxplots. The $y = 0$ line is given by the black dashed line.

graphical or saturated covariance structure are unbiased. The CMEVM predictions exhibit a small positive bias, although the reason for this is unclear.

S3.6.3 Multivariate t-distribution

For this study, $\boldsymbol{X}$ follows a MVT distribution with mean $\boldsymbol{\mu} = \mathbf{0}$, $k = 5$ degrees of freedom, and a dispersion matrix with inverse consistent with $\mathcal{G}$ in Section S3.3.1. We consider weak positive, strong positive, and negative associations in $\boldsymbol{X}$ in Sections S3.6.3, S3.6.3, and S3.6.1, respectively. For all simulations, the dependence threshold u_{Y_i} is set to the 0.8-quantile of the standard Laplace distribution for all $i \in V$, resulting in approximately $1,000$ excesses per conditioning variable. As with the previous examples, parameter estimates are omitted unless of specific interest.

Weak positive dependence

In this simulation, we ensure the associations between components are weakly positive i.e., entries in the dispersion matrix are strictly positive but less than 0.17. For prediction, we

Figure S3.20: Boxplots of empirical and model-based estimates of $\Gamma_{|i}$, for each $i \in V$, when the data is generated from a MVT distribution with weak positive associations. Each row corresponds to the conditioning variable i and each column corresponds to the correlation parameter. The colour of the boxplots distinguishes the different models. Black dashed lines show $y = 0$.

set $u_{X_i} = 0.75$ for each $i \in V$.

Figure S3.20 shows empirical and model-based estimates of the conditional precision matrix $\Gamma_{|i}$. The estimated structure of the conditional precision matrix from both the graphical and saturated SCMEVMs is again consistent with the empirical version. Analysis of other parameter estimates has been omitted, other than noting that estimates of $\beta_{j|i}$ from the one-step SCMVEM are generally larger than corresponding estimates from the two- and three-step SCMEVMs.

Figure S3.21 (left panel) shows the bias in the conditional survival curves of $X_j \mid X_3 > u_{X_3}$ for $j \in V_{|3}$. Similar to the MVG and MVL examples, the SCMVEM with graphical covariance structure is unbiased for all curves, and the EHM exhibits positive bias for low values of u_{X_j} that decreases as u_{X_j} increases. We have also included the estimated curve from the CMEVM to show that there is little difference between the estimates from the CMEVM and

the SCMEVM with a graphical covariance structure.

Figure S3.21 (right panel) shows the bias in $\mathbb{P}[\boldsymbol{X}_{|3} > u_{\boldsymbol{X}_{|3}} \mid X_3 > u_{X_3}]$. As expected, the CMEVM estimates are unbiased, the EHM exhibits positive bias, and the three-step SCMEVM with independent residuals exhibits negative bias because not all the components of $\boldsymbol{X}_{|3}$ are independent given X_3. Interestingly, the fully parametric SCMEVMs with graphical or saturated covariance structures exhibit a very small negative bias. Despite this, the three-step SCMEVM with graphical structure is the least biased and variable model as it minimises the MAE and RMSE for 38 and 46 of the 75 conditional tail probabilities, respectively. For comparison, the CMEVM minimises the metrics 21 and 13 times, respectively. This suggests there is little difference between the CMEVM and the SCMEVM in this scenario.

Assessing diagnostic plots from the SCMEVM raises no concerns about the model fit. Therefore, to fix the slight underestimation from the SCMEVMs in Figure S3.21 (right panel), we may need to increase N used in Algorithm 3.6 of the main text (we let $N = 250,000$ in this simulation). Alternatively, we may wish to simulate data from the fitted model for $\boldsymbol{X} \mid X_i > u_{X_i}$ for each $i \in V$, rather than using the method outlined in Section 3.3 in the main text which simulates data for the entire domain, with the extreme region corresponding to at least one component being extreme.

Strong positive dependence

We now allow strong positive associations between the components. The only note on the parameter estimates is that the CMEVM dependence parameters tend to be higher for the one-step SCMEVM compared to the two- and three-stepwise SCMEVMs, and the marginal AGG parameters (excluding the shape) tend to be lower for the one-step SCMEVM. Setting $u_{X_i} = 1.25$ for each $i \in V$, Figure S3.22 (left panel) shows the bias in the conditional survivor curves of $X_j \mid X_2 > u_{X_2}$ for $j \in V_{|2}$, and for the EHM and the three-step SCMEVM with graphical covariance structure. The EHM is biased for low values of u_{X_j}, but this diminishes as u_{X_j} increases. In contrast, the three-step SCMEVM with a graphical covariance structure is unbiased across all curves. Figure S3.22 (right panel) shows the bias in $\mathbb{P}[\boldsymbol{X}_{|3} > u_{\boldsymbol{X}_{|3}} \mid X_3 > u_{X_3}]$. The EHM has positive bias, while the CMEVM and the stepwise SCMEVMs with graphical or saturated covariance structures are unbiased. Once again, the SCMEVMs with graphical covariance structure are the least biased and the least variable, minimising both the MAE and RMSE for 48 of the 75 conditional tail probabilities.

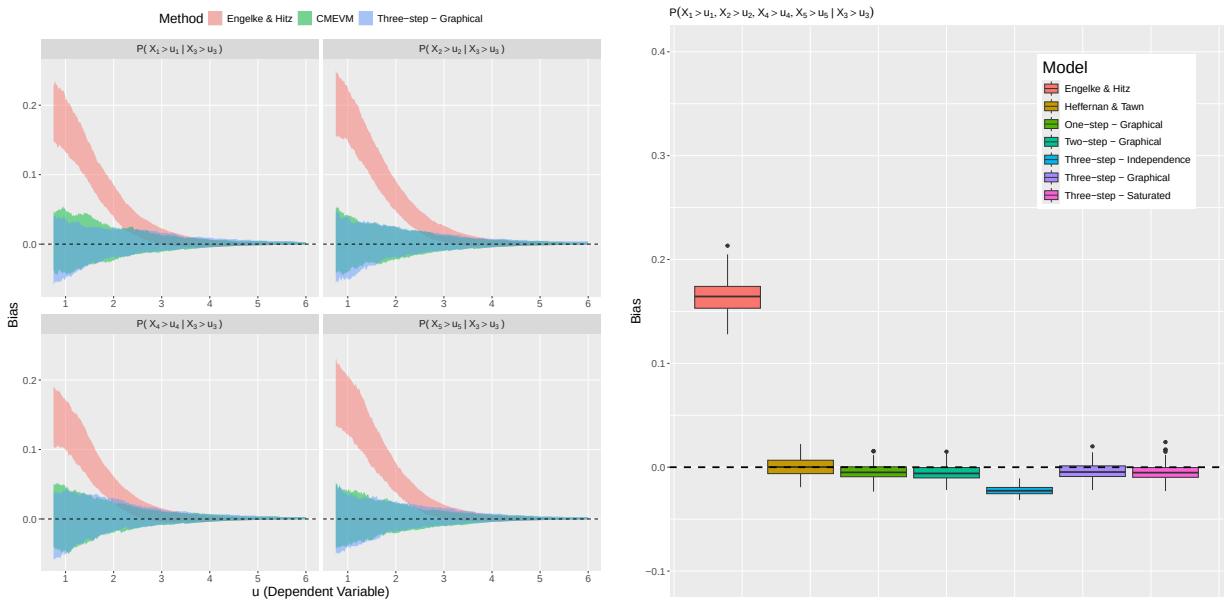

Figure S3.21: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j > u_{X_j} \mid X_3 > u_{X_3}]$, for each $j \in V_{|3}$, where $\boldsymbol{X}$ follows a MVT distribution with weak positive associations (left). The bias from the EHM, the CMEVM, and the three-step SCMEVM, assuming a graphical covariance structure for the residuals are in pink, green, and blue, respectively. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|3} > u_{\boldsymbol{X}_{|3}} \mid X_3 > u_{X_3}]$ (right). The fill of the boxplots distinguishes the different models. Black dashed lines show $y = 0$.

Negative dependence

Finally, we allow for weak negative associations between the components. For prediction, we set u_{X_i} to the 0.9-quantile from a dataset of size 10^6 simulated from the true distribution for each $i \in V$. Figure S3.23 (left panel) shows 95% confidence intervals for the conditional cumulative distribution curves of $X_j \mid X_3 > u_{X_3}$ from the EHM and the three-step SCMEVM with graphical covariance structure for $j \in V_{|3}$. As with the MVG and MVL distributions with negative associations, the three-step SCMEVM captures all curves perfectly, whereas the EHM always underestimates. Figure S3.23 (right panel) shows the bias in $\mathbb{P}[\boldsymbol{X}_{|5} < u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$. The EHM performs poorly due to its inability to capture the negative dependence, while predictions from the CMEVM and the SCMEVMs with graphical or saturated covariance structures are unbiased. The SCMEVMs with a graphical covariance structure minimise the MAE and RMSE metrics for 79% and 72% of all the conditional tail probabilities, respectively.

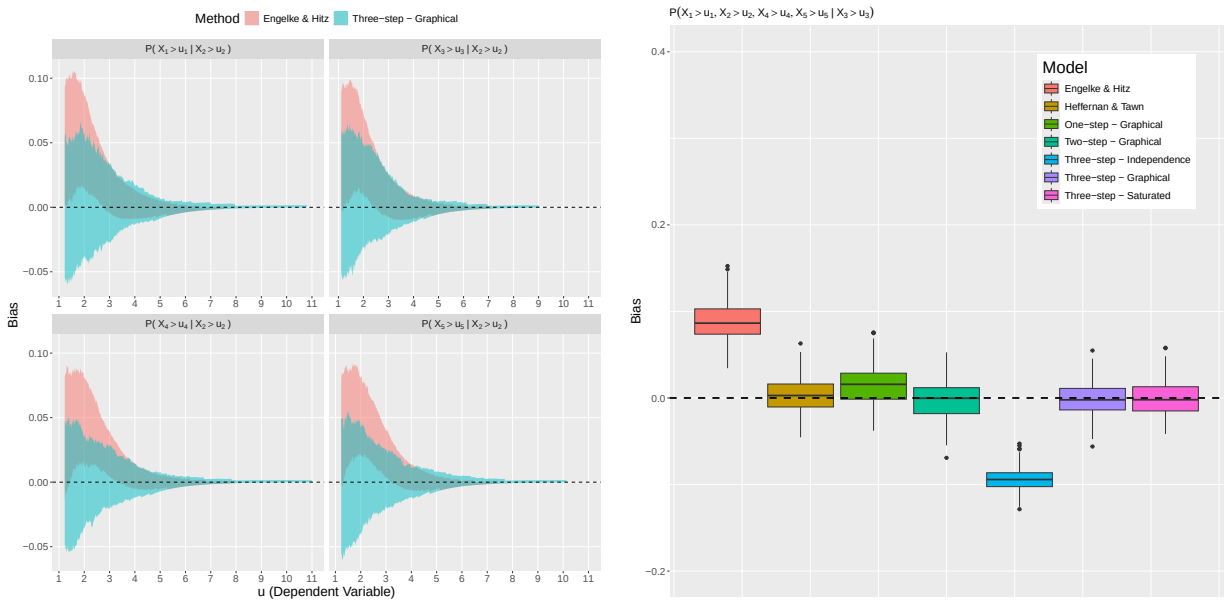

Figure S3.22: Polygon plots detailing 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j > u_{X_j} \mid X_2 > u_{X_2}]$ for $j \in V_{|2}$, where X follows a MVT distribution with strong positive associations (left). The bias from the EHM and the three-step SCMEVM, assuming a graphical covariance structure for the residuals are in pink and blue, respectively. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|3} > u_{\boldsymbol{X}_{|3}} \mid X_3 > u_{X_3}]$ (right). The bias from the various models is denoted by the fill of the boxplots. Black dashed lines show $y = 0$.

S3.6.4 Multivariate Pareto distribution

To test the SCMEVM under AD, we repeat the simulation study for $\boldsymbol{X}$ with a MVP distribution such that the parameter matrix is consistent with $\mathcal{G}$ in Section S3.3. For the CMEVM and SCMEVMs, we only use data above the dependence threshold u_{Y_i} set at the 0.90-quantile of the standard Laplace distribution for all $i \in V$. The EHM uses all data since, by construction, the data is on standard Pareto margins. For prediction, $u_{X_i} = 11$ for each $i \in V$.

To illustrate the difference between the one-, two-, and three-step parameter estimation procedures, Figure S3.24 shows boxplots of MLEs of the dependence and AGG parameters. Here, we would expect $\hat{\alpha}_{j|i} = 1$ and $\hat{\beta}_{j|i} = 0$ since the data are asymptotically dependent. While $\hat{\alpha}_{j|i} \approx 1$ for all the methods, the variability from the one-step method is much greater. This is linked to the variability in the AGG parameters, particularly the location parameter, being larger under the one-step method. Although some MLEs of $\beta_{j|i} > 0$ under the two- and three-step methods, these are closer to 0 than the corresponding estimates under the one-step method. This highlights that the one-step method cannot guarantee that the first-order extremal dependence is restricted to the CMEVM dependence parameters, and why

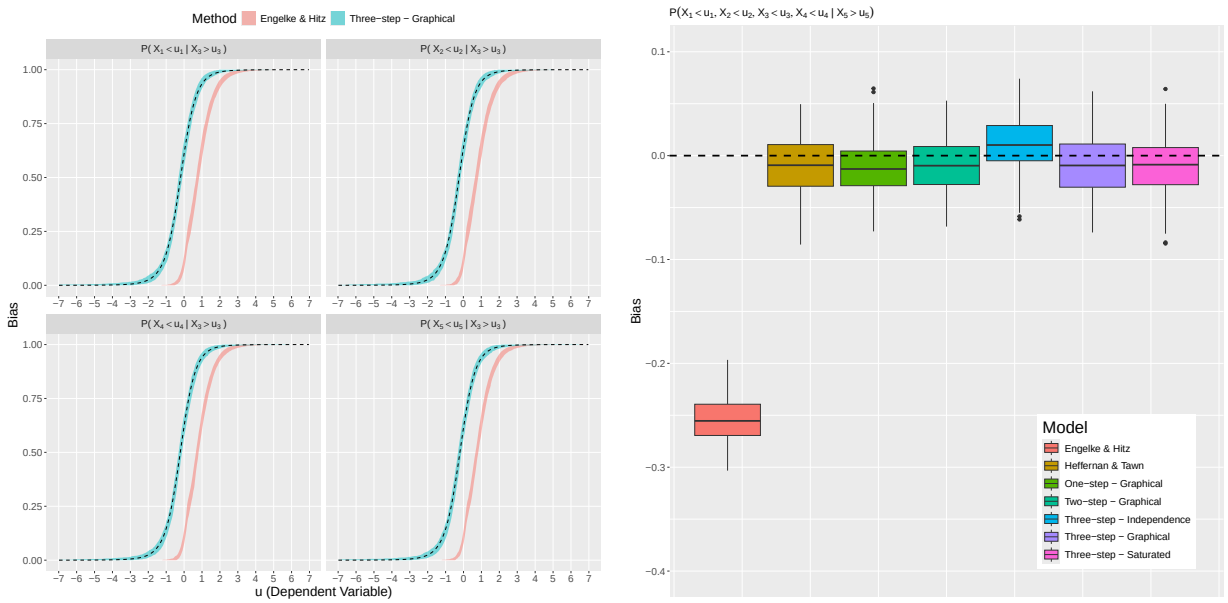

Figure S3.23: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j < u_{X_j} \mid X_3 > u_{X_3}]$ for $j \in V_{|3}$, where $\boldsymbol{X}$ follows a MVT distribution with negative associations (left). The bias from the EHM and the three-step SCMEVM, assuming a graphical covariance structure for the residuals are in pink and blue, respectively. The true conditional cumulative distribution curves are given by the black dashed lines. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|5} < u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$ (right). The bias from the various models is denoted by the fill of the boxplots. The $y = 0$ line is given by the black dashed line.

we prefer the three-step estimation method.

Figure S3.25 shows empirical and SCMEVM model-based estimates of $\Gamma_{|i}$ and transformed model-based estimates from the EHM. Again, the empirical structure is retained in all models. The EHM and SCMEVMs have a very close correspondence, although the former has less variability due to the larger sample size. Figure S3.26 (left panel) shows the bias in the conditional survivor curves of $X_j \mid X_5 > u_{X_5}$ for $j \in V_{|5}$. Both the EHM and the three-step SCMEVM with graphical covariance are unbiased, but estimates from the former are slightly less variable due to the larger sample size. Figure S3.26 (right panel) shows the bias in $\mathbb{P}[\boldsymbol{X}_{|5} > u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$. As expected, the EHM is unbiased while the SCMEVMs exhibit a slight negative bias, perhaps due to the α parameter not converging to the true value, $\alpha = 1$, that lies on the edge of the parameter space. However, the bias in the SCMEVMs with graphical and saturated covariances is small. Considering all 75 conditional tail probabilities, the EHM minimises both the MAE and RMSE metrics the majority of the time. Excluding the EHM, which is the only model specifically designed for AD data, the SCMEVMs with graphical covariance minimise the metrics 83% and 89% of the time, respectively, suggesting that the SCMEVM is an acceptable alternative to the EHM when the extremal dependence

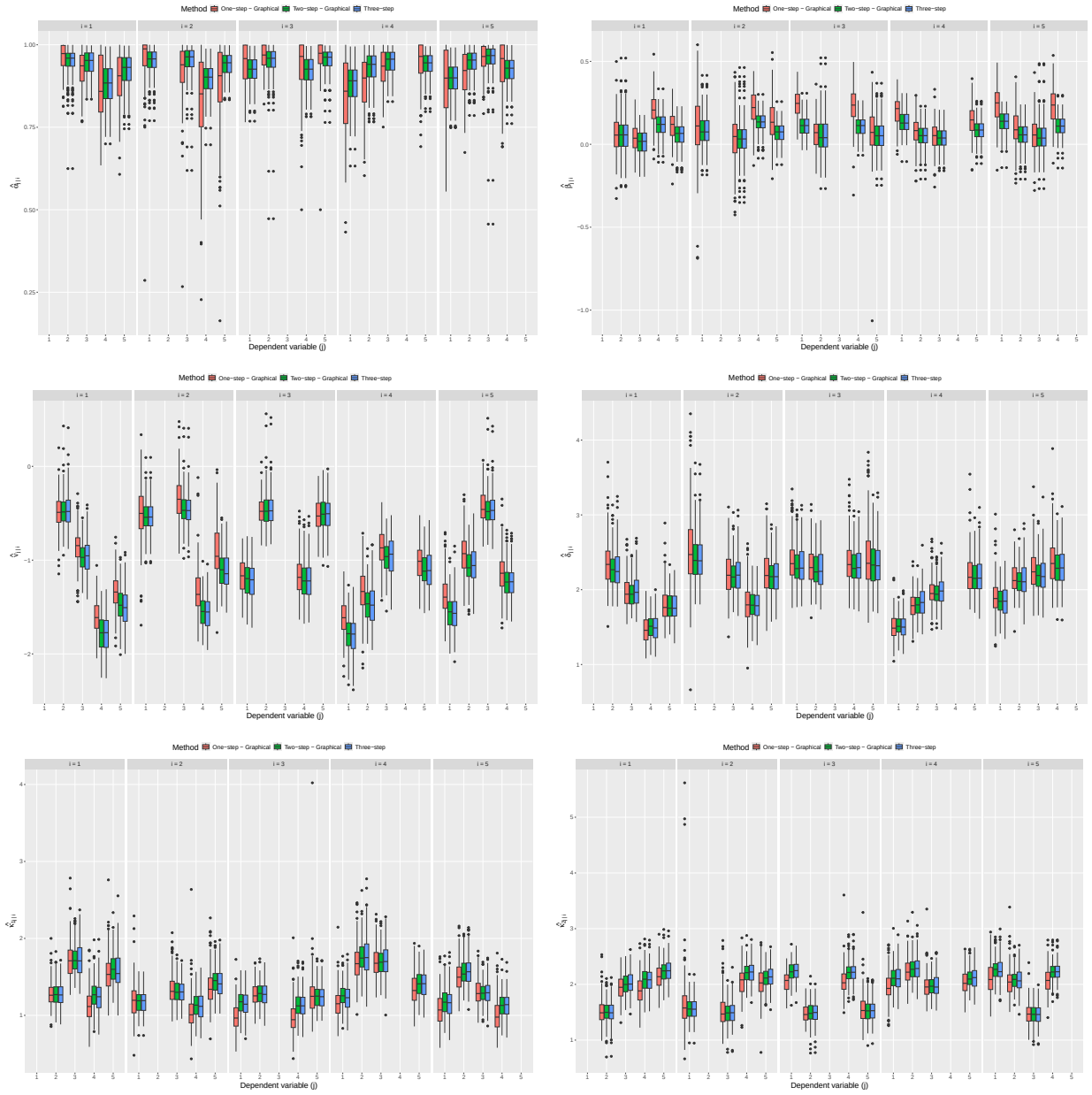

Figure S3.24: Boxplots of MLEs for $\alpha_{j|i}$ (top left), $\beta_{j|i}$ (top right), $\nu_{j|i}$ (centre left), $\delta_{j|i}$ (centre right), $\kappa_{1_{j|i}}$ (bottom left), and $\kappa_{2_{j|i}}$ (bottom right) for distinct $i, j \in V$. Each column corresponds to the conditioning variable i. The different models are denoted by the fill of the boxplots.

class cannot be pre-determined.

Figure S3.25: Boxplots of empirical and model-based estimates of $\Gamma_{|i}$, for each $i \in V$, when the data is generated from a MVP distribution. Each row corresponds to the conditioning variable i, and each column corresponds to the correlation parameter. The various models are denoted by the colour of the boxplots. Black dashed lines show $y = 0$.

S3.7 Application to the upper Danube River basin

We first present additional figures for the model comparison made in Section 5 of the main text. We then use EGlearn (Engelke et al., 2025) to learn the graphical structure for the upper Danube River basin under the assumption of asymptotic dependence. Using the learnt structure, we then assess the predictive qualities of the model.

S3.7.1 Additional figures for Section 5

We obtain 200 non-parametric bootstrap samples of the declustered river discharge data from the upper Daube River basin. For each bootstrapped dataset, we fit: (i) the EHM and (ii) the three-step SCMEVM, both with a graphical covariance structure given by the undirected tree induced by the flow connections of the upper Danube River basin (Figure 1, left panel, of the main text); (iii) the three-step SCMEVM with saturated covariance structure; (iv) the three-step SCMEVM with a graphical covariance structure inferred from the data (Figure 6,

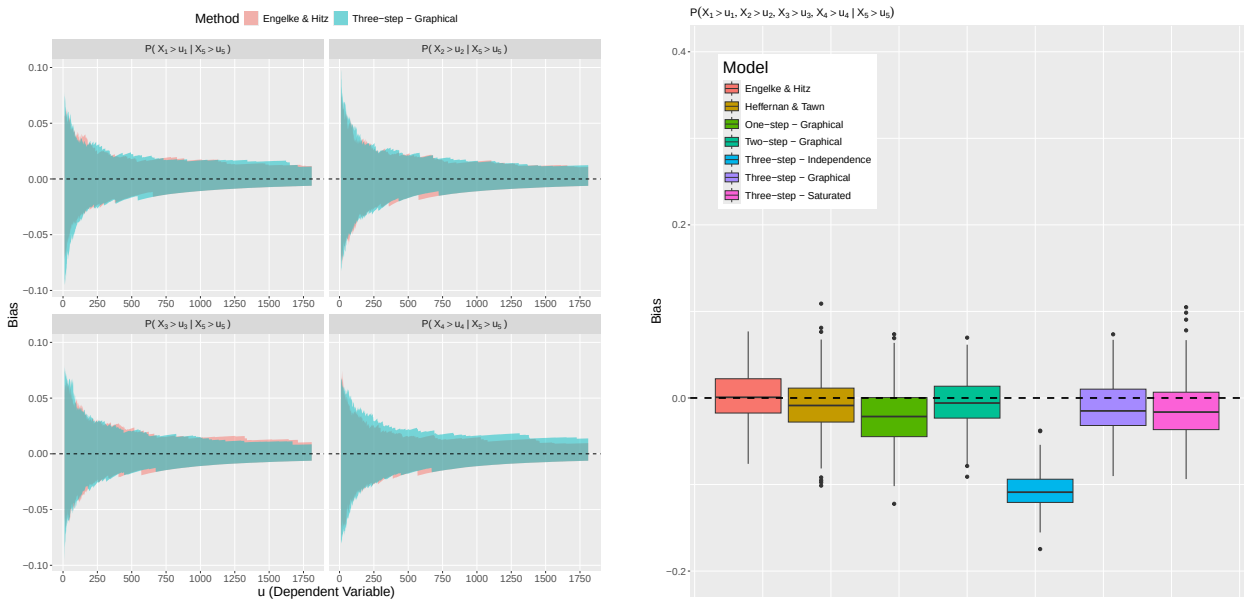

Figure S3.26: Polygon plots detailing pointwise 95% confidence intervals, over 200 samples, of the bias in $\mathbb{P}[X_j > u_{X_j} \mid X_5 > u_{X_5}]$ for $j \in V_{|5}$, where $\boldsymbol{X}$ follows a MVP distribution (left). The bias from the EHM and the three-step SCMEVM, assuming a graphical covariance structure for the residuals are in pink and blue, respectively. Boxplots of the bias in $\mathbb{P}[\boldsymbol{X}_{|5} > u_{\boldsymbol{X}_{|5}} \mid X_5 > u_{X_5}]$ (right). The bias from the various models is denoted by the fill of the boxplots. Black dashed lines show $y = 0$.

right panel, of the main text).

For each fitted model and bootstrapped dataset, we obtain a single simulation, which is used for prediction. Empirical and model-based estimated for $\chi_{i,j}(u)$ are obtained for $i, j \in V$, $i > j$, and $u \in \{0.8, 0.85, 0.9\}$, where $V = \{1, \ldots, 31\}$. The point estimates in Figure S3.27 are the median estimates over the two sets of estimates for $\chi_{i,j}(u)$. As in Figure 7 of the main text, the SCMEVMs better capture the extremal dependence structure than the EHM. Figure S3.28 shows a similar comparison but for $\chi_A(u)$ where $A \subset V$ are 500 randomly sampled triplets, and $u \in \{0.8, 0.85, 0.9\}$. Again, the SCMEVMs better capture the extremal dependence in the upper Danube River basin. Although for higher thresholds the EHM is less biased for moderately dependent triplets, the bias for low dependent triplets increases as the threshold increases. Similar conclusions can be made for (iv), however, the magnitude of the bias is much smaller. Model (iv) appears to perform better than (ii) and rectifies the systemic underestimation in (iii). Thus, (iv) appears to have the best predictive performance across multiple components.

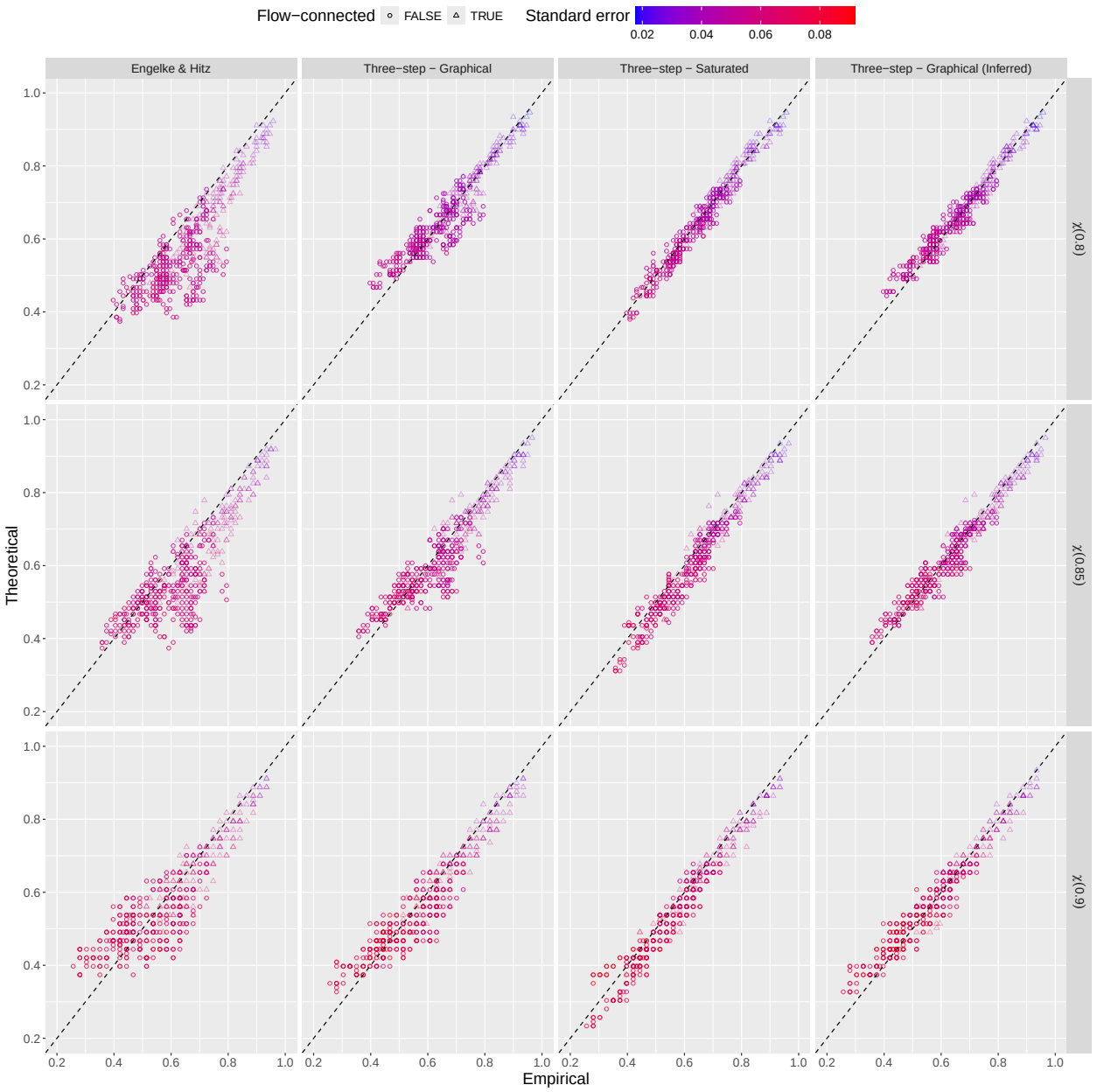

Figure S3.27: Empirical and model-based estimates of $\chi_{i,j}(u)$ for $u \in \{0.8, 0.85, 0.9\}$ (top to bottom), and $i, j \in V$ but $i > j$. Model-based estimates use the EHM (left) and the three-step SCMEVM with graphical covariance (centre left), with structure given in Figure 1 (left panel) of the main text, the three-step SCMEVM with saturated covariance (centre right) and graphical covariance (right) with structure given in Figure 6 (right panel) of the main text. Black dashed lines show $y = x$. Circles (triangles) show flow-connected (flow-unconnected). The colour shows the standard error of the model-based estimates.

S3.7.2 Comparison with EGlearn

In their seminal paper, Engelke and Hitz (2020) focus on learning block graphical structures. Since then, the literature on learning the graphical structure for data that is assumed to

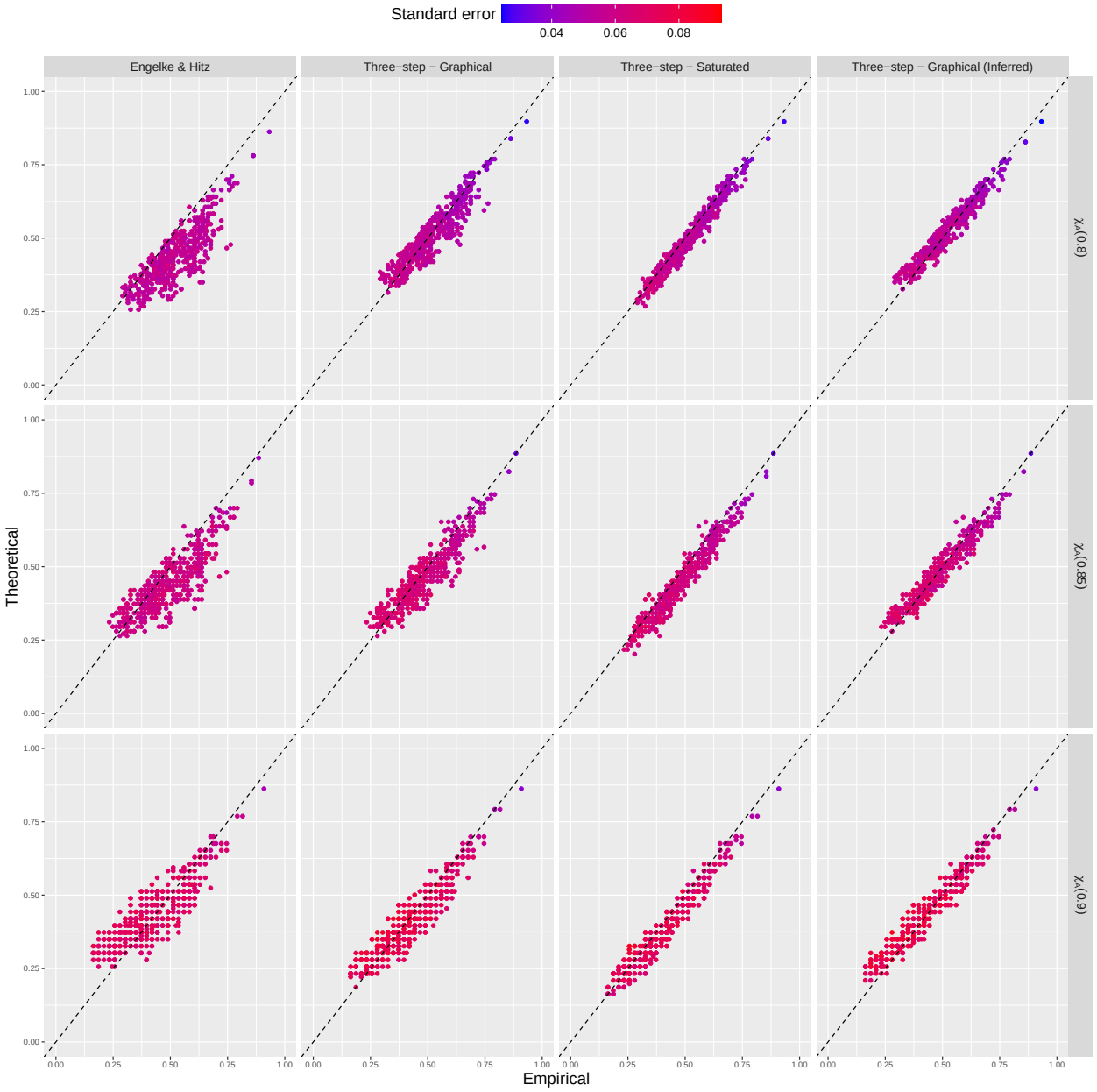

Figure S3.28: Empirical and model-based estimates of $\chi_A(u)$ for $u \in \{0.8, 0.85, 0.9\}$ (top to bottom) for 500 randomly selected triplets of $A \subset V$. Model-based estimates use the EHM (left) and the three-step SCMEVM with graphical covariance (centre left), both with structure given in Figure 1 (left panel) of the main text, the three-step SCMEVM with saturated covariance (centre right) and graphical covariance (right) with structure given in Figure 6 (right panel) of the main text. Black dashed lines show $y = x$. The colour shows the standard error of the model-based estimates.

follow a Hüsler-Reiss distribution has exploded; see Engelke et al. (2024a) for a thorough review. One such method is EGlearn (Engelke et al., 2025). Figure S3.29 compares the inferred graphical structures for the upper Danube River basin using EGlearn with model

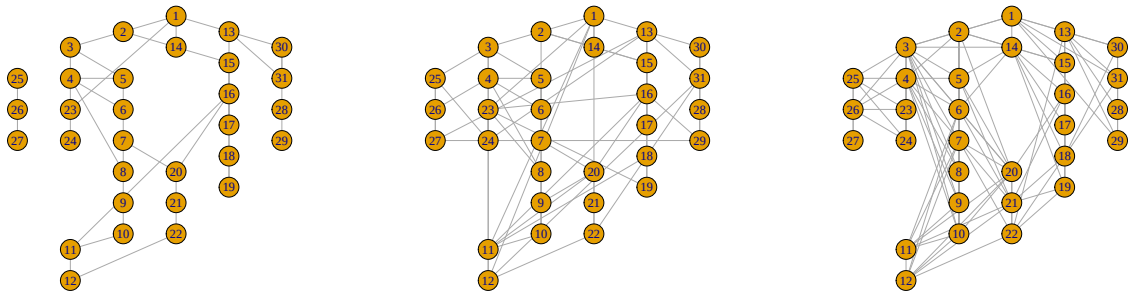

Figure S3.29: Inferred graphical structure of the upper Danube River basin using EGlearn, where the model selection criterion is MBIC (left) and AIC (centre), and the method proposed in the main text (right).

selection criterion MBIC (left) and AIC (centre), and the method proposed in the main text (right).

Using EGlean with MBIC results in a very sparse graph. The inferred graph has 42 edges compared to 30 edges in the graph induced by the flow connections in the river structure. However, the inferred graph is disconnected. This is likely because stations 23-27 are asymptotically independent due to their latitude (north of the main Danube River) (Engelke et al., 2025). Thus, Engelke et al. (2025) remove these stations from their analysis as they do not satisfy the assumption underlying the model. Since we wish to make a like-for-like comparison between EGlearn and our method when we predict from the model, we therefore use the inferred graph using AIC. This graph is much denser (71 edges in total), with clear connections between the tributaries due to shared rainfall events that are not captured by the graph induced by the river flow. This compares to 127 edges in the graph using our method. While there are 56 more edges, the connected structures in the centre and right panel of Figure S3.29 are not dissimilar, as they both: capture the underlying structure of the river; connect geographically close but disconnected tributaries; and create additional connections along existing tributaries.

Using the same data and methods as in Section S3.7.1, we fit the EHM to the AIC EGlearn-inferred graphical structure. For comparison, we fit the SCMEVM with the same graphical structure. Furthermore, we fit the SCMEVM to the inferred graphical structure obtained with our method.

For each fitted model and for each bootstrapped dataset, we obtain a single simulation, which is used for prediction. Empirical and model-based estimated for $\eta_{i,j}(u)$ are obtained for $i, j \in V$, $i > j$, and $u \in \{0.8, 0.85, 0.9\}$, where $V = \{1, \ldots, 31\}$. The point estimates

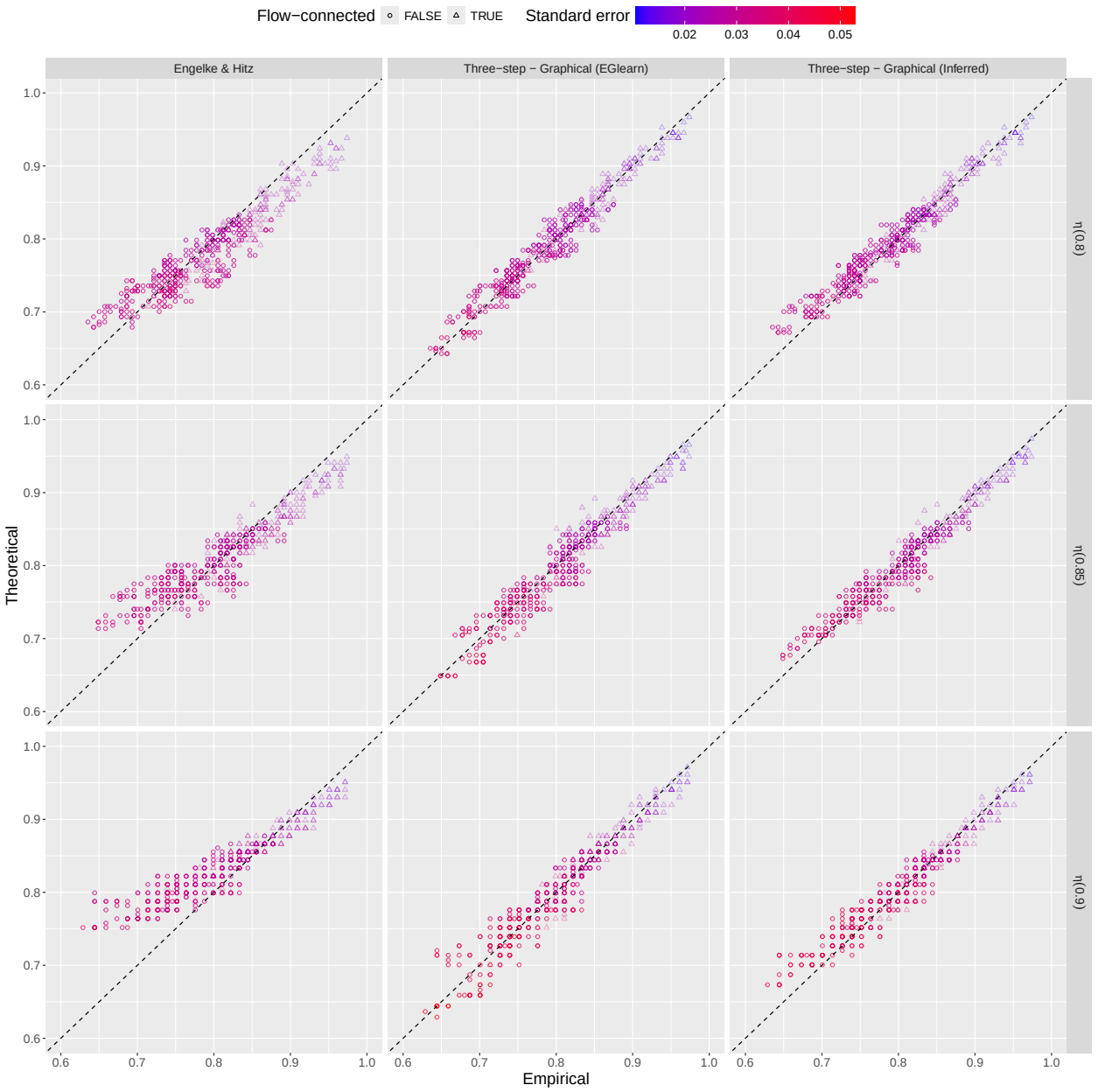

Figure S3.30: Empirical and model-based estimates of $\eta_{i,j}(u)$ for $u \in \{0.8, 0.85, 0.9\}$ (top to bottom), and $i, j \in V$, but $i > j$. Model-based estimates use the EHM (left) and the three-step SCMEVM with graphical covariance (centre), both with structure given in Figure S3.29 (centre panel), and the three-step SCMEVM graphical covariance (right), with structure given in Figure S3.29 (right panel). Black dashed lines show $y = x$. Circles (triangles) show flow-connected (flow-unconnected). The colour shows the standard error of the model-based estimates.

in Figure S3.30 are the median estimates over the two sets of estimates for $\eta_{i,j}(u)$. The left panels show predictions from the EHM are improved compared to using the graph induced by the flow connections, with closer alignment to the $y = x$ line for $u < 0.9$ and less variability

across all values of u. However, there are deviations in the left tail where the dependence is weaker and closer to complete independence, particularly as u tends towards 1. The majority of deviations can be attributed to when one or more of the sites north of the main Danube River (23 -27) is included in the pair, supporting the claim that these stations tend to exhibit AI with other stations (Engelke et al., 2025). Since the EHM assumes complete AD, it is unsurprising that the model overestimates in these cases. The deviation is resolved by using the three-step SCMEVM (centre panels). In addition, using the SCMEVM reduces the bias for pairs of sites with strong positive dependence. For a like-for-like comparison of the methods, we compare the left and right panels, where the EHM is fit using the EGlearn-inferred graph and the three-step SCMEVM is fit using the SCMEVM-inferred graph. The SCMEVM exhibits lower variability, has closer agreement to the $y = x$ line, and resolves the overestimation of the left-tail in the EHM, highlighting the need for a flexible model that can capture all extremal dependence classes. Interestingly, the SCMEVM performs better in the left tail with the EGlearn-inferred graph than the SCMEVM-inferred graph, suggesting the proposed method for learning the graph may require refinement.

Bibliography

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023 – 2050.

Blanchet, J. and Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, 5(3):1699–1725.

Bolin, D., Simas, A. B., and Wallin, J. (2024). Gaussian Whittle–Matérn fields on metric graphs. *Bernoulli*, 30(2):1611 – 1639.

Casey, A. and Papastathopoulos, I. (2023). Decomposable tail graphical models. *arXiv preprint arXiv:2302.05182*.

Debusho, L. K. and Diriba, T. A. (2021). Conditional modelling approach to multivariate extreme value distributions: application to extreme rainfall events in South Africa. *Environmental and Ecological Statistics*, 28(3):469–501.

Engelke, S., Hentschel, M., Lalancette, M., and Röttger, F. (2024a). Graphical models for multivariate extremes. *arXiv preprint arXiv:2402.02187*.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932.

Engelke, S., Hitz, A. S., Gnecco, N., and Hentschel, M. (2024b). *graphicalExtremes: Statistical Methodology for Graphical Extreme Value Models*. R package version 0.3.2.

Engelke, S., Lalancette, M., and Volgushev, S. (2025). Learning extremal graphical structures in high dimensions. *arXiv preprint arXiv:2111.00840*.

Engelke, S., Malinowski, A., Kabluchko, Z., and Schlather, M. (2015). Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):239–265.

Ferreira, A. and de Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717 – 1737.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2019). *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. R package version 1.11.

Gouldby, B., Méndez, F., Guanche, Y., Rueda, A., and Mínguez, R. (2014). A methodology for deriving extreme nearshore sea conditions for structural design and flood risk analysis. *Coastal Engineering*, 88:15–26.

Gouldby, B., Wyncoll, D., Panzeri, M., Franklin, M., Hunt, T., Hames, D., Tozer, N., Hawkes, P., Dornbusch, U., and Pullen, T. (2017). Multivariate extreme value modelling of sea conditions around the coast of England. *Proceedings of the Institution of Civil Engineers - Maritime Engineering*, 170(1):3–20.

Heffernan, J. E. and Resnick, S. I. (2007). Limit laws for random vectors with an extreme component. *The Annals of Applied Probability*, 17(2):537 – 571.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Keef, C., Papastathopoulos, I., and Tawn, J. A. (2013). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, 115:396–404.

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.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436 – 1462.

Murphy, C., Tawn, J. A., and Varty, Z. (2025). Automated threshold selection and associated inference uncertainty for univariate extremes. *Technometrics*, 67(2):215–224.

Nacereddine, N. and Goumeidane, A. B. (2019). Asymmetric generalized Gaussian distribution parameters estimation based on maximum likelihood, moments and entropy. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 343–350.

Nagler, T. and Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89.

Neal, J., Keef, C., Bates, P., Beven, K., and Leedal, D. (2013). Probabilistic flood risk mapping including spatial dependence. *Hydrological Processes*, 27(9):1349–1363.

Oesting, M. and Stein, A. (2018). Spatial modeling of drought events using max-stable processes. *Stochastic environmental research and risk assessment*, 32:63–81.

Papastathopoulos, I. (2016). Conditional independence and conditioned limit laws. *Statistics & Probability Letters*, 112:1–4.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramos, A. and Ledford, A. (2011). An alternative point process framework for modeling multivariate extreme values. *Communications in Statistics-Theory and Methods*, 40(12):2205–2224.

Richards, J., Tawn, J. A., and Brown, S. (2022). Modelling extremes of spatial aggregates of precipitation using conditional methods. *The Annals of Applied Statistics*, 16(4):2693 – 2713.

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

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131.

Ross, E., Astrup, O. C., Bitner-Gregersen, E., Bunn, N., Feld, G., Gouldby, B., Huseby, A., Liu, Y., Randell, D., Vanem, E., and Jonathan, P. (2020). On environmental contours for marine and coastal design. *Ocean Engineering*, 195:106194.

Röttger, F., Coons, J. I., and Grosdos, A. (2023). Parametric and nonparametric symmetries in graphical models for extremes. *arXiv preprint arXiv:2306.00703*.

Shewchuk, J. R. et al. (1994). An introduction to the conjugate gradient method without the agonizing pain. *Carnegie-Mellon University. Department of Computer Science Pittsburgh*.

Shooter, R., Ross, E., Ribal, A., Young, I. R., and Jonathan, P. (2021). Spatial dependence of extreme seas in the North East Atlantic from satellite altimeter measurements. *Environmetrics*, 32(4):e2674.

Shooter, R., Ross, E., Tawn, J., and Jonathan, P. (2019). On spatial conditional extremes for ocean storm severity. *Environmetrics*, 30(6):e2562.

Simpson, E. S. and Wadsworth, J. L. (2021). Conditional modelling of spatio-temporal extremes for Red Sea surface temperatures. *Spatial Statistics*, 41:100482.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231.

Smith, R. L. (1990). Max-stable processes and spatial extremes. *Unpublished manuscript*, pages 1–32.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *The Annals of Statistics*, 14(1):138 – 150.

Stephenson, A. G., Shaby, B. A., Reich, B. J., and Sullivan, A. L. (2015). Estimating spatially varying severity thresholds of a forest fire danger rating system using max-stable extreme-event modeling. *Journal of Applied Meteorology and Climatology*, 54(2):395–407.

Towe, R., Tawn, J., Lamb, R., and Sherlock, C. (2019). Model-based inference of conditional extreme value distributions with hydrological applications. *Environmetrics*, 30(8):e2575.

Wadsworth, J. and Tawn, J. (2013). A new representation for multivariate tail probabilities. *Bernoulli*, 19(5B):2689 – 2714.

Wadsworth, J. and Tawn, J. (2022). Higher-dimensional spatial extremes via single-site conditioning. *Spatial Statistics*, 51:100677.

Wan, P. and Zhou, C. (2025). Graphical lasso for extremes. *arXiv preprint arXiv:2307.15004*.

Westra, S. and Sisson, S. A. (2011). Detection of non-stationarity in precipitation extremes using a max-stable process model. *Journal of Hydrology*, 406(1-2):119–128.

Winter, H. C. and Tawn, J. A. (2016). Modelling heatwaves in central France: a case-study in extremal dependence. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 65(3):345–365.

Winter, H. C. and Tawn, J. A. (2017). kth-order Markov extremal models for assessing heatwave risks. *Extremes*, 20:393–415.

Yildirim, I. (2012). Bayesian inference: Gibbs sampling. *Technical Note, University of Rochester*, 3:4.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Chapter 4

Conditional Extremes with Metric Graphs

Conditional Extremes with Metric Graphs

Abstract

Multivariate extreme value theory aims to model the dependence structure of random vectors to quantify the probability and magnitude of joint extreme events. Numerous methods exist for modelling joint extreme events of processes in a continuous spatial setting. However, modelling joint extreme events of stochastic processes on non-Euclidean spaces, such as road and river networks, has received little attention. Where it has, multivariate, rather than spatial, methods are generally employed due to the complex spatial relationships exhibited by the process. Consequently, the models are unable to make predictions at unobserved locations. Asadi et al. (2015) overcome this by modelling declustered river discharges in the upper Danube River basin using a max-stable Brown-Resnick process. However, the model assumes full asymptotic dependence, which is unrealistic over large spatial domains. To address this limitation, we extend the conditional multivariate extreme value model (Heffernan and Tawn, 2004) to model stochastic processes on non-Euclidean spaces, the first attempt to our knowledge, by modelling the residual process with a Gaussian Whittle-Matérn field on a metric graph (Bolin et al., 2024). We apply our model to the declustered river discharges in the upper Danube River basin and obtain predictions over the entire river network.

4.1 Introduction

Flooding is a natural phenomenon which poses one of the largest risks to communities globally. Fluvial flooding regularly causes costly mass destruction, and even casualties when the flooding is severe. Thus, accurate prediction of flooding is vital for government authorities to help them assess the most effective flood mitigation policies and manage resources and emergency response units when floods occur. Given the risks of flooding and the benefits of accurate prediction, the literature for statistical modelling of river flows is extensive. However, river flows have complex spatial and temporal dependence structures that are difficult to capture. Consequently, the preferred approach is to model precipitation using a geospatial statistical model and feed simulations from this model into a rainfall-runoff model to obtain simulated river flows (Wheater et al., 2005). The rainfall-runoff model may be either probabilistic or deterministic, but in either case, detailed knowledge of catchment characteristics is required. More recently, both statistical (Asadi et al., 2015) and machine learning (Luppichini et al., 2024) methods have been implemented to model spatial flow events. Our objective is to develop an interpretable statistical model with low computational cost that can be used to forecast flood risk at any location on a river network.

Our proposed solution combines two distinct methodologies: extreme value analysis and geostatistical models for network data. In recent years, extreme value analysis has leveraged many methods from geostatistics to develop spatial models for extremes. Davison et al. (2012) and Huser and Wadsworth (2022) provide thorough reviews of these approaches, which include max-stable processes (Schlather, 2002), $r-$Pareto processes (de Fondeville and Davison, 2018), Gaussian scale mixture models (Huser et al., 2017), and the spatial conditional extreme value model (Wadsworth and Tawn, 2022), among others. Despite the array of extreme spatial models available, river flow at multiple gauging stations is generally treated as a multivariate process within the extremes literature because it is defined on a non-Euclidean space. Consequently, these models can only provide predictions at observed locations, which are often sparsely located over the network, conveniently located for data collection, and may not correspond to regions of interest for practitioners, i.e. they may not represent where the most extreme events occur. Our solution addresses this limitation by applying a particular class of specialised Gaussian processes.

First, we illustrate the pitfalls of applying a model based on Euclidean distance to extreme river flows. Define the process $\{X(s) : s \in \mathcal{S}\}$ for some spatial domain $\mathcal{S} \subset \mathbb{R}^2$ with $\boldsymbol{x}_t = (x_t(s_1), \ldots, x_t(s_d))$ for $t = 1, \ldots, n$, denoting independent and identically distributed (IID) realisations of the process $\{X(s)\}$ at d sampling locations $\boldsymbol{s} = (s_1, \ldots, s_d) \subset \mathcal{S}$ and n time points. In our example, we have $n = 428$ realisations of declustered river flows at $d = 31$ gauging stations in the upper Danube River basin (Asadi et al., 2015) made available in the `graphicalExtremes` package (Engelke et al., 2025) in `R` (R Core Team, 2025). For this dataset, Figure 4.1 shows pairwise estimates of the coefficient of tail dependence (Coles et al., 1999) $\{\eta_q(s_1, s_2) : s_1, s_2 \in \mathcal{S}, q = 0.9)\}$ as a function of distance between the sites for three distance metrics, $d(s_1, s_2) = ||s_1, s_2||$.

For Euclidean distance (left panel), the estimates do not decay smoothly. Further, some of the flow-connected pairs have stronger extremal dependence than some flow-unconnected pairs despite having a larger Euclidean distance. In geostatistical modelling, for stationary covariance structures, the correlation between sites generally decreases as the distance between them increases (Diggle and Ribeiro, 2007). For spatial extremes with stationary covariance structures, the extremal dependence between sites should decrease as the distance between them increases, meaning sites with a small Euclidean separation distance exhibit asymptotic dependence (AD), and those further away exhibit asymptotic independence (AI) (Huser and Wadsworth, 2022). For this dataset, one could argue that the covariance structure is not stationary, as the dependence downstream is unlikely to be the same as the dependence upstream. However, if we falsely assume a stationary covariance structure, the more pertinent issue is that the distance metric is inappropriate since it ignores the inherent

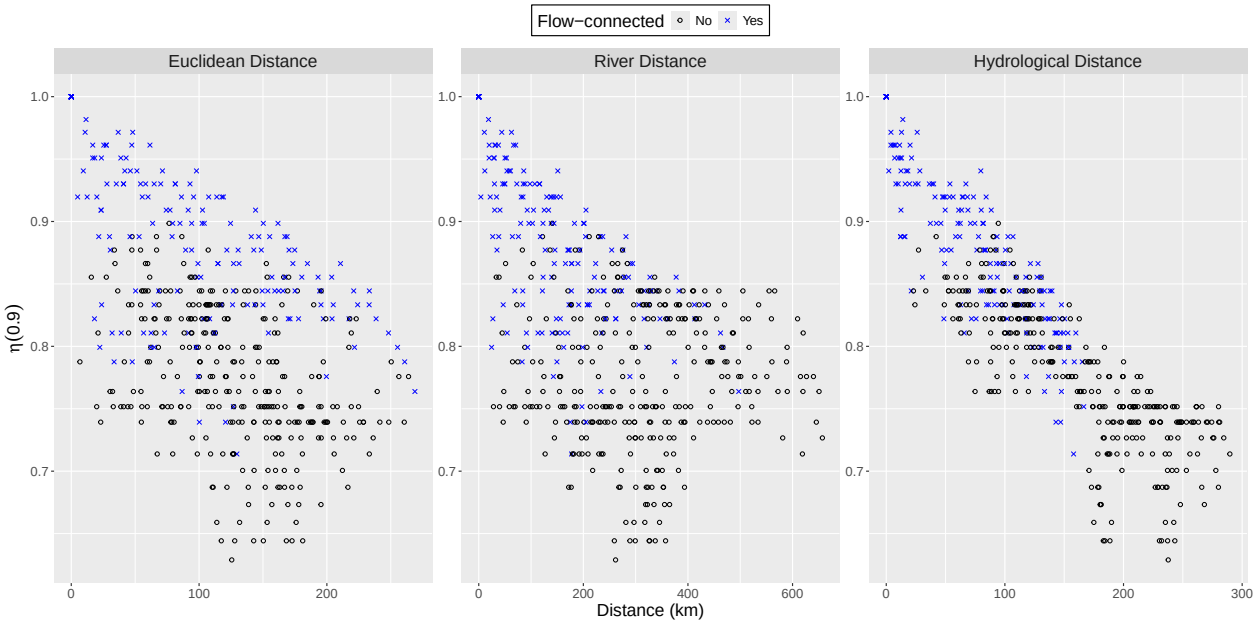

Figure 4.1: Pairwise estimates of the coefficient of tail dependence against Euclidean (left), river (centre), and hydrological (right) distance. Flow-connected and flow-unconnected pairs are denoted by blue crosses and black circles, respectively.

structure in the data. By using river distance (centre panel), the decay is smoother, as the structure of the data is included in the distance metric. This is even further resolved by using hydrological distance (right panel). Note, river distance refers to distance along the river network, and hydrological distance is the distance between the altitude-weighted centroids of the sub-catchments of the gauging stations (see Asadi et al. (2015) for details).

Despite the intuitive results obtained using these different distances, many of the standard correlation functions used in geostatistical modelling require a Euclidean distance metric to guarantee a valid (positive definite) spatial dependence structure (Ver Hoef et al., 2006). For example, for the river distance metric, we do not obtain a valid spatial dependence structure when we employ the Matérn correlation function

$$\rho(d(s_1, s_2)) = \rho(|h|) = \frac{\tau^{-2}}{2^{\nu-1}\Gamma(\nu + 1/2)(4\pi)^{1/2}\kappa^{2\nu}}(\kappa|h|)^{\nu}K_{\nu}(\kappa|h|), \tag{4.1.1}$$

where $|h| = d(s_1, s_2)$, Γ is the standard gamma function, $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind of order ν, and the parameters $\tau \in \mathbb{R}$, $\kappa > 0$, and $\nu \in (0, 1/2]$ control the variance, correlation range, and spatial smoothness, respectively. While modelling stochastic processes that are valid on networks is challenging, several developments in this area have been made, including the work of Cressie et al. (2006), Ver Hoef and Peterson (2010), O'Donnell et al. (2014), and Bolin et al. (2024).

The remainder of this chapter is structured as follows. Section 4.2 overviews currently available methodology for modelling extreme river flows, while Section 4.3 details our proposed approach, including methods for inference and prediction. We apply the proposed model to the declustered river discharges in the upper Danube River basin (Asadi et al., 2015) in Section 4.4 before concluding with a discussion of the limitations of the proposed methodology and how to alleviate them, as well as potential alternative modelling approaches, in Section 4.5.

4.2 Background

The conditional multivariate extreme value model (CMEVM) (Heffernan and Tawn, 2004) is a popular choice for modelling multivariate extremes as it can capture AD when $\alpha_{s|s_0} = 1$ and $\beta_{s|s_0} = 0$, complete independence when $\alpha_{s|s_0} = 0$, and AI for all other parameter combinations (see Section 4.3.2 for technical details). Note, the CMEVM cannot capture both AD and AI concurrently. For example, if two random variables, X_1 and X_2, on common margins experience their extreme events both simultaneously and independently, then the model cannot capture this. To the best of our knowledge, the only framework that can capture such behaviour is geometric extremes (Nolde and Wadsworth, 2022).

Since the geometric extremes framework is still in its infancy, the CMEVM remains a popular choice for modelling. Conditional on an extreme event at one location, the distribution of the remaining locations can be characterised. Keef et al. (2009) extend the model to account for the temporal lag in extreme river flows between locations at different parts of the network. A further extension handles missing observations, which occur due to gauge failures and non-overlapping data collection periods at different gauging stations. They achieve this by infilling the data using a multivariate Gaussian distribution conditional on the observed values. Towe et al. (2019) note this infilling technique is computationally expensive and fails to rectify the underlying curse of dimensionality problem with the original CMEVM. They therefore propose combining non-parametric kernel density estimators with a Gaussian copula to handle missing observations. Irrespective of the infilling technique, both models are multivariate and are unable to predict river flow at unobserved locations. To obtain river flow over the entire river network, Lamb et al. (2010) spatially interpolate the simulations of Keef et al. (2009). However, the interpolation uses Euclidean distance between catchment centroids (similar to hydrological distance), which may not result in a smooth spatial map that necessarily reflects river flow, particularly at confluence points, or highly localised flood events.

A secondary concern with applying a multivariate model is the dimensionality. For instance, the original CMEVM fails in high dimensions (Farrell et al., 2024) and models based on geometric extremes can currently only handle [edit: moderate dimensions] (the framework has been extended to 8- and 10-dimensional random vectors by Murphy-Barltrop et al. (2024) and Monte et al. (2025), respectively). Typically, spatial processes have many more sampling locations than this. A possible solution is dimension reduction, which can be achieved using graphical models (Engelke and Hitz, 2020; Farrell et al., 2024), clustering (Rohrbeck and Tawn, 2021) and extremal principal component analysis (PCA) (Rohrbeck and Cooley, 2023). Note that clustering and extremal PCA are somewhat adjacent to using regionalised models in hydrological applications (Bergström, 1991), where the latter models are defined using hydrologically similar catchment areas based on knowledge of the river basins, while the former uses regions that are determined empirically.

Graphical models allow us to consider higher-dimensional problems by leveraging conditional independence structures to reduce the number of dependence parameters in the model. In the multivariate Gaussian case, this is equivalent to having a sparse precision matrix. The declustered river flows from the upper Danube River basin have been modelled using multivariate extreme value graphical models by Engelke and Hitz (2020), who use a multivariate Pareto distribution, and Farrell et al. (2024), who use an extension of the CMEVM. The former model assumes full AD, while the latter captures both AD and AI. However, neither can simulate river flow at unobserved locations on the river network without post-inference interpolation (see, for example, Lamb et al. (2010)).

Spatial clustering requires identification of $M < d$ clusters, such that pairs of stations within the same cluster have stronger dependence than pairs of stations in different clusters. Thus, if we assume the strength of dependence between pairs of stations is only dependent on which cluster they belong to, then the dimensionality of the problem is vastly reduced to M. Rohrbeck and Tawn (2021) employ a reversible jump Markov chain Monte Carlo algorithm that determines the clusters, and marginal and dependence parameters jointly. When applied to winter river flows in northern England, this resulted in three clusters for 45 gauging stations, which drastically reduces the number of parameters in the model. Rohrbeck and Cooley (2023) alternatively use extremal PCA to reduce the dimensionality of the same dataset. The model identifies 12 principal components, which, although not as large a reduction in the parameter space, is still a reduction. While the dimensionality reduction is helpful, the models are multivariate in nature and unable to make predictions at unobserved locations.

Brunner et al. (2019) can make predictions of river flow at unobserved locations by mod-

elling the dependence between gauging stations using a Fisher copula. The Fisher copula is non-parametric and relies on the empirical correlation matrix. The authors extend the empirical correlation matrix to include correlations between unobserved locations using a fitted correlogram. The fitted correlogram is obtained by modelling the empirical correlation matrix as a function of a distance metric using either parametric functions or splines. Thus, the model includes correlations for unobserved locations and can thereby simulate river flows at said locations. Similar results could be obtained by interpolating simulations from the unextended model. However, the fitted correlogram is likely to be more reliable, provided there is no directional effect, for unobserved locations that are not on a river segment between two observed gauges.

Moving to spatial extreme models, Jóhannesson et al. (2022) use a latent Gaussian modelling framework to model river flows over the United Kingdom. They model annual maxima river flow at 554 gauging stations using a generalised extreme value (GEV) distribution, incorporating both covariates and independent spatial random effects into the GEV parameters. The spatial random effects are modelled via a Gaussian Markov Random Field (GMRF) on a finite mesh (Lindgren et al., 2011). The model diagnostics and simulations appear reasonable, however, the GMRF approximates a continuous-space random field rather than one on a network. Further flows are assumed to be independent conditional on the marginal parameters. Hence, it is unclear if the model provides cohesive predictions for stations on the same river.

The closest solution to our method is provided by Asadi et al. (2015), who also models river discharges from the upper Danube River basin using a Brown-Resnick process. They include information from the river in a similar manner to the variance component model of Ver Hoef and Peterson (2010); they define the kernel for the Brown-Resnick process as the sum of a kernel for all stations based on "hydrological" distance and a kernel for flow-connected stations using river distance. However, the modelling strategy has several limitations. First, a Brown-Resnick process assumes full AD, which is not valid for this dataset (Farrell et al., 2024), meaning predictions are likely to overestimate the river flow between stations where there is AI. Secondly, despite the model fitting the data well, a digital elevation model is required to derive most of the covariates at unobserved locations, making it difficult to extend the model to further unobserved locations without considerable data processing. Thirdly, while the model can be used to make predictions over the entire river basin, Asadi et al. (2015) only provide such predictions for return levels estimated using the marginal model. Finally, their marginal model uses the non-homogeneous Poisson point process (NHPPP) representation of the peaks over threshold model, meaning the predictions are only valid above a high threshold. Given that the dependence model assumes full AD, all the predictions

will be jointly extreme, meaning localised flooding events cannot be examined using this model.

As we have discussed, existing methods have multiple limitations. Our solution, detailed in Section 4.3, takes inspiration from the methods proposed by Keef et al. (2009), Lamb et al. (2010), and Asadi et al. (2015), and takes advantage of the very recently developed class of Gaussian Whittle-Matérn fields for metric graphs (Bolin et al., 2024), to overcome all the limitations of the existing models.

4.3 Methodology

In this section, we present models for both the marginal behaviour and the dependence structure of extreme river flow events. Our approach follows the standard routine for a multivariate or spatial extreme value analysis: first model the marginal behaviour, use the resulting model to standardise the margins, and then model the dependence structure. Throughout, we work with log-transformed data $\{\log(X(s))\}$ to ensure that predictions are always positive. Note that we assume that the process is stationary in time. This is not a restrictive assumption since many methods for event identification exist (see Section 4.5 for more details), and the declustered river discharges in the upper Danube River basin have been pre-processed to remove temporal dependence (see Asadi et al. (2015) for more details).

4.3.1 Marginal Modelling

Our first step is to model the marginal distribution. Recall that $\{X(s)\}$ is a spatial process, which we assume to be stationary in time but not in space. Making predictions at unobserved locations requires estimation of the marginal distribution at locations for which there is no data. Thus, we seek relationships between the marginal model parameters and spatial covariates. One way to achieve this is via a regionalised GEV derived from a NHPPP (Asadi et al., 2015). A regionalised approach is common in hydrological applications (Fischer and Schumann, 2021; Merz and Blöschl, 2005) as it encourages parsimony by leveraging information from "hydrologically-close" stations. Although the model is well-fitting, the covariates required for prediction are not readily available at unobserved locations without access to a digital elevation model.

We instead model the spatial trends using generalised additive models (GAMs) for extreme value distributions (Youngman, 2022). For each location $s \in \mathcal{S}$, assume that there exists an associated vector of covariates $\boldsymbol{C}(s)$. We now assume that, for any location, the observations can be modelled as independent random variables which follow a non-stationary GEV, with

cumulative distribution function (CDF)

$$\mathbb{P}[\log(X(s)) \leq x \mid \boldsymbol{C}(s) = \boldsymbol{c}(s)] = \exp\left\{-\left[1 + \xi(\boldsymbol{c}(s))\left(\frac{x - \upsilon(\boldsymbol{c}(s))}{\psi(\boldsymbol{c}(s))}\right)\right]^{-1/\xi(\boldsymbol{c}(s))}\right\}, \quad (4.3.1)$$

where $\upsilon(\cdot)$, $\psi(\cdot)$, and $\xi(\cdot)$ are the covariate-dependent location, scale and shape parameters, respectively.

To capture spatial trends using a GAM, each of the GEV parameters is expressed as a sum of basis functions. Specifically, $\upsilon(\boldsymbol{c}(s)) = \eta_\upsilon(\boldsymbol{c}(s))$, $\log(\psi(\boldsymbol{c}(s))) = \eta_\psi(\boldsymbol{c}(s))$, and $\xi(\boldsymbol{c}(s)) = \eta_\xi(\boldsymbol{c}(s))$, such that

$$\eta_*(\boldsymbol{c}(s)) = \gamma_0^* + \sum_{k=1}^{K}\sum_{d=1}^{D_k}\gamma_{kd}^* g_{kd}^*(\boldsymbol{c}(s)), \quad (4.3.2)$$

where γ_0^* is the intercept term for parameter $*$, K is the number of covariates, D_k is the basis dimension of covariate k, and γ_{kd}^* and g_{kd}^* are basis coefficients and functions for parameter $*$, respectively. The log link function is used to ensure $\psi(\boldsymbol{c}(s)) \in \mathbb{R}_+$. Models in equation (4.3.1) can be fitted using the `EVGAM` package (Youngman, 2025) in `R` (R Core Team, 2025). If equation (4.3.1) is maximised directly, the model would likely overfit to the data as the splines in equation (4.3.2) will be too flexible. Instead, a penalty term that includes predefined smoothing parameters is added to the likelihood function derived from equation (4.3.1) to avoid under-smoothing the splines. To ensure the splines are flexible, but not too flexible, the likelihood function is maximised using standard restricted maximum likelihood (REML) techniques (Wood, 2017).

Note that many spline families exist to set up the basis functions g_{kd}. The default in `EVGAM` are thin-plate regression splines, which, compared to alternatives, have fewer coefficients relative to the number of data points. However, they do not have "knots" in the traditional sense, so user information on where changes in the function are likely to occur, such as a confluence point in a river, cannot be incorporated into the basis functions. If such information were vital, one could use piecewise splines that are polynomial between each of the specified knots. However, piecewise polynomial splines can overfit if the degree of the polynomial is too large. Thus, the splines should be carefully chosen to reflect the user's modelling goals.

4.3.2 Dependence Modelling

For the dependence structure, we utilise the CMEVM. As discussed in Section 4.2, the CMEVM has previously been used to model the dependence of river flows between gauging stations (Farrell et al., 2024; Keef et al., 2009; Towe et al., 2019). However, by using multivari-

ate rather than spatial models, predictions at unobserved locations can only be obtained by subsequent interpolation (Lamb et al., 2010). While the CMEVM has been extended to the spatial case to account for processes on Euclidean spaces (Wadsworth and Tawn, 2022), the model has yet to be extended to processes on non-Euclidean spaces, such as river networks. Lastly, all variants of the CMEVM require that the data be on common margins. Without loss of generality, it is usual to transform $\{\log(X(s))\}$ onto standard Laplace margins $\{Y(s)\}$ (Keef et al., 2013) using a double application of the probability integral transform (PIT).

The spatial CMEVM is an approximation to the asymptotic behaviour of the process conditional on observing an extreme event at one location. Formally, let $s_0 \in \mathcal{S}$ denote the conditioning site and suppose that $Y(s_0) > u^Y_{s_0}$ for some high threshold $u^Y_{s_0}$. We assume there exists normalising functions $a_{s|s_0}(\cdot)$ and $b_{s|s_0}(\cdot)$, such that as $u^Y_{s_0} \to \infty$,

$$\left(\left\{ \frac{Y(s) - a_{s|s_0}(Y(s_0))}{b_{s|s_0}(Y(s_0))} : s \in \mathcal{S} \right\}, \; Y(s_0) - u^Y_{s_0} \right) \Bigg| \; Y(s_0) - u^Y_{s_0} \xrightarrow{d} \left(\{Z_{s_0}(s) : s \in \mathcal{S}\}, \; E \right), \tag{4.3.3}$$

where the residual process $\{Z_{s_0}(s) : s \in \mathcal{S}\}$ is non-degenerate for all $s \in \mathcal{S} \backslash s_0$, in the sense that the process places no mass on $+\infty$, and is constrained such that $Z_{s_0}(s_0) = 0$ almost surely. Furthermore, in the limit, the residual process is independent of the conditioning component $Y(s_0)$ and has marginal mean and variance parameters denoted by $\mu_{s|s_0}$ and $\sigma^2_{s|s_0}$, respectively. The form of the statistical model follows by assuming that expression (4.3.3) holds for finite but high values of $u^Y_{s_0}$. Independence means that inference can be undertaken separately on: $Y(s_0) - u^Y_{s_0} \mid Y(s_0) > u^Y_{s_0}$, the normalising functions, and the residual process $\{Z_{s_0}(s) : s \in \mathcal{S}\}$. The first is trivially a standard exponential distribution E, but more care is needed in modelling the normalising functions and residual process.

Normalising functions

For any pair of sites $(s, s_0) \in \mathcal{S}$, we follow the original CMEVM and assume that the normalising functions in equation (4.3.3) take the form

$$a_{s|s_0}(y(s_0)) = \alpha_{s|s_0} y(s_0) \text{ and } b_{s|s_0}(y(s_0)) = y(s_0)^{\beta_{s|s_0}},$$

where $\alpha_{s|s_0} \in [-1, 1]$ and $\beta_{s|s_0} \in (-\infty, 1]$ are unknown parameters to be estimated. Note, we will restrict our $\alpha_{s|s_0} \in (0, 1]$ for reasons explained in the next two paragraphs. Further, we assume that $\alpha_{s_0|s_0} = 1$ and $\beta_{s_0|s_0} = 0$. These parameters, and the mean $\mu_{s|s_0}$ and variance $\sigma^2_{s|s_0}$ parameters, vary with respect to the conditioning and dependent locations. Consequently,

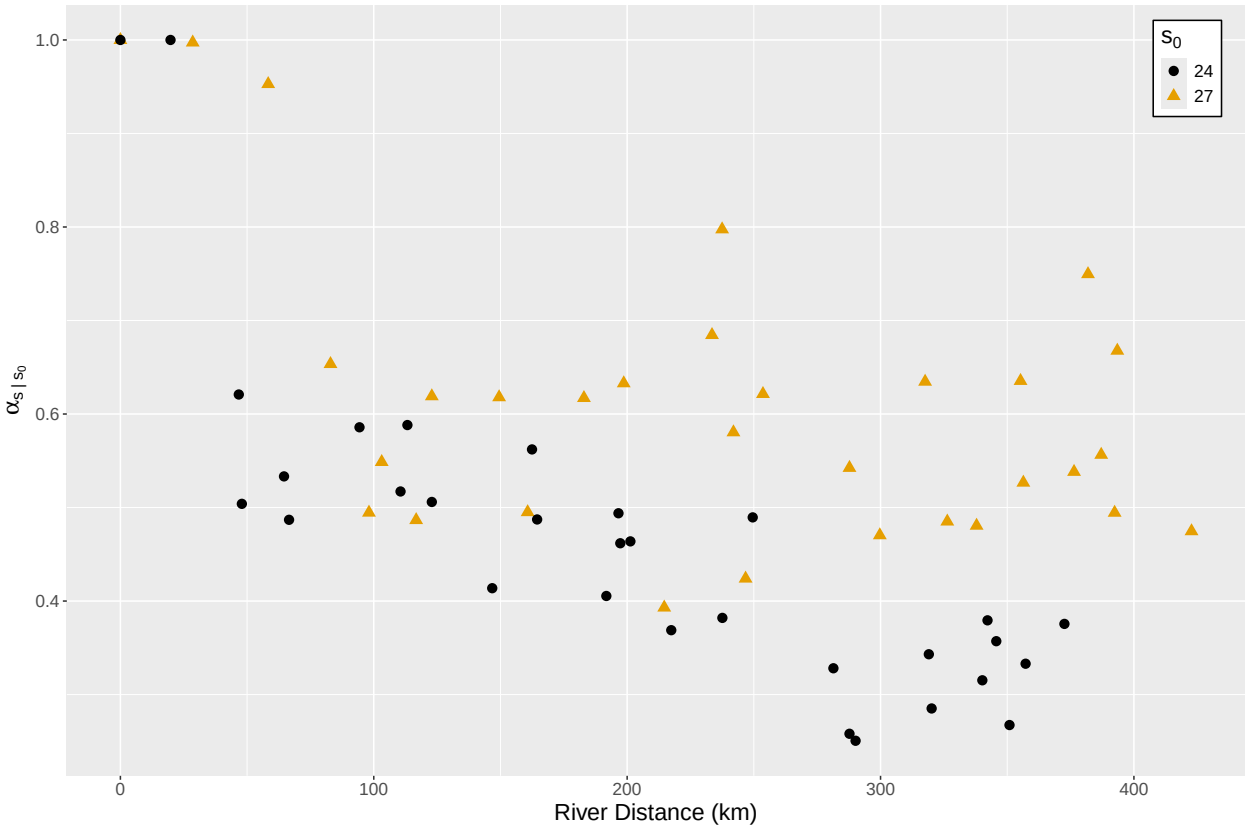

Figure 4.2: Scatter plot comparing estimates of $\alpha_{s|s_0}$ from the original CMEVM against river distance from the conditioning station. The estimates when we condition on stations 24 and 27 are represented by the black circles and orange triangles, respectively.

they can only be estimated for locations at which data is observed. To make predictions for all $s \in \mathcal{S}$, we need to interpolate these parameters across all locations.

In the continuous spatial setting, this is achieved by modelling the parameters $\alpha_{s|s_0}$ and $\beta_{s|s_0}$ as parametric functions of the distance between s and s_0 (Richards et al., 2022; Simpson and Wadsworth, 2021), parametric functions that account for both this distance and direction (Shooter et al., 2021), and semi-parametric functions of distance (Simpson et al., 2023). Finding parametric functions is more difficult when the data is observed on non-Euclidean spaces. To illustrate this, Figure 4.2 shows pairwise estimates of $\alpha_{s|s_0}$ against river distance from the conditioning location when we condition on stations 24 and 27. A comparison between only two conditioning locations shows that the functional form of $\alpha_{s|s_0}$ differs considerably between conditioning sites. Further, while conditioning on site 24 demonstrates a clear trend with distance, conditioning on site 27 does not. The reason for this is that dependence is determined not just by distance along the river, but also by the proximity of catchments and the connectedness of the two sites.

To address this, we model the CMEVM parameters using GAMs with spatial covariates. Rather than directly estimating the GAM coefficients by incorporating this structure in equation (4.3.3), we found that a two-step approach was more reliable. For clarity, we found that estimating all the CMEVM parameters ($\alpha_{s|s_0}$, $\beta_{s|s_0}$, $\mu_{s|s_0}$, $\sigma^2_{s|s_0}$) simultaneously caused $\beta_{s|s_0}$ to be poorly estimated. This links to the issues in the original CMEVM that the residual distribution can explain properties related to the extremal dependence structure, which is undesirable. Therefore, we elect for a two-step approach.

First, we obtain pairwise maximum likelihood estimates (MLEs) using the original CMEVM. We denote these estimates $\hat{*}^{HT}_{s|s_0}$ for parameter $*$. Note that this assumes the margins of the residuals are symmetric, despite them often exhibiting asymmetry (Farrell et al., 2024; Ross et al., 2018). Although it would be preferable to fit a model with this feature embedded, as discussed above, $\beta_{s|s_0}$ was poorly estimated when we assumed symmetric margins and therefore extending the model for asymmetric margins posed further fitting issues.

Second, we fit GAMs to the pairwise estimates as follows. By assuming that $\alpha_{s|s_0} \in (0, 1]$, we can use a logistic GAM. The restriction on the parameter space implies that all pairs of sites have positive extremal dependence, which is a plausible assumption given that spatial surfaces have been constructed using a multivariate event identification procedure (see Section 4.5). For the remaining parameters, we fit a GAM with Gaussian errors to $\log(1 - \hat{\beta}^{HT}_{s|s_0})$, $\hat{\mu}^{HT}_{s|s_0}$ and $\log(\hat{\sigma}^{2HT}_{s|s_0})$, where the transformations are selected to preserve the CMEVM parameter spaces. For example, since $\hat{\beta}^{HT}_{s|s_0} \in (-\infty, 1)$ then $\log(1 - \hat{\beta}^{HT}_{s|s_0}) \in (-\infty, \infty)$. We then obtain estimates $\hat{*}^{GAM}_{s|s_0}(\boldsymbol{c}(s))$ for parameter $*$ using the `gam` function in `R` (R Core Team, 2025), where $\boldsymbol{c}(s)$ is a vector of covariates at $s \in \mathcal{S}$. For consistency with the marginal modelling approach in Section 4.3.1, we only consider thin-plate regression splines and optimise the smoothing parameters using REML. For clarity, this means we fit a GAM to the d data points $\hat{*}^{HT}_{s|s_0}$ for each conditioning location $s_0 \in \mathcal{S}$.

Residual process

We first define the standardised residual process $\{W_{s_0}(s) : s \in \mathcal{S}\}$ as the pointwise transformation of $\{Z_{s_0}(s) : s \in \mathcal{S}\}$ given by

$$W_{s_0}(s) = \frac{Z_{s_0}(s) - \mu^{GAM}_{s|s_0}(\boldsymbol{c}(s))}{\sigma^{GAM}_{s|s_0}(\boldsymbol{c}(s))}. \tag{4.3.4}$$

Observe that $W_{s_0}(s_0) = 0$ with probability 1, as we assume that $a_{s|s_0}(y(s_0)) = y(s_0)$ and $b_{s|s_0}(y(s_0)) = 0$ at $s_0 \in \mathcal{S}$. We assume that $\{W_{s_0}(s)\}$ can be modelled as a Gaussian Whittle-Matérn field on a metric graph (Bolin et al., 2024). To define this class of Gaussian fields we

start with a metric graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which is a collection of vertices $\mathcal{V}$ and edges $\mathcal{E}$ such that the vertex set $\mathcal{V} = \boldsymbol{v} = (v_1, \ldots, v_m)$ consists of the end-points of every edge in the graph. In the case of a river network, the vertices correspond to sources and junctions that determine the graphical structure of the physical network. These locations, which we term the original locations, are demonstrated for the upper Danube River basin in Figure 4.3. In addition, a metric graph $\mathcal{G}$ ***must*** be associated with a distance metric that allows a shortest path to be defined between any pair of locations on the graph. On a river network, river distance provides this metric. The Gaussian Whittle-Matérn field on $\mathcal{G}$ is then defined as the solution u to the stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta_{\mathcal{G}})^{\delta/2}(\tau u) = \mathcal{W}, \tag{4.3.5}$$

where $\Delta_{\mathcal{G}}$ is the Kirchoff-Laplacian and $\mathcal{W}$ is a Gaussian white noise process on $\mathcal{G}$. The parameters τ, $\kappa > 0$, and $\delta = \nu + 1/2$ are analogous to those in the Matérn correlation function in equation (4.1.1); they control the variance, practical correlation range, and sample path regularity, respectively. The practical correlation range controls the rate at which the correlation function decays to 0 and is measured in the same way as in standard geostatistical models (see Diggle and Ribeiro (2007) for more details), with an alternative distance metric, namely the distance along the graph. The sample path regularity controls the differentiability of the sample path. For example, $\delta = 1$ means the path is continuous, while $\delta = 2$ ensures continuity and differentiability.

To facilitate the use of the Gaussian Matérn-Whittle field for statistical inference, Bolin et al. (2024) focus on the case $\delta \in \mathbb{N}$, which results in the field having a Markov structure (or GMRF) in which the edges are conditionally independent given the nodes. This conditional independence permits a bridge representation of the process along the edges. This property can be leveraged for computationally efficient inference by allowing the deconstruction of a high-dimensional multivariate Gaussian likelihood into the product of two much smaller-dimensional Gaussian likelihoods. Bolin et al. (2024) give two alternative inference approaches based on this construction. In the first (direct) method, the data are assumed to be a direct draw from the field described above. In the second (indirect) method, the data are assumed to be observed with measurement error. These errors are assumed to be independent and identically distributed over space. Although we do not have measurement error, the indirect method provides a better model fit (see Section 4.4.2).

To construct the associated likelihood Bolin et al. (2024), use an extended metric graph $\bar{\mathcal{G}} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$, by incorporating the observation locations as vertices into the graph. The vertex set $\bar{\mathcal{V}}$ consists of two distinct sets $\boldsymbol{v} = (v_1, \ldots, v_m)$ and $\boldsymbol{s} = (s_1, \ldots, s_d)$ that correspond to

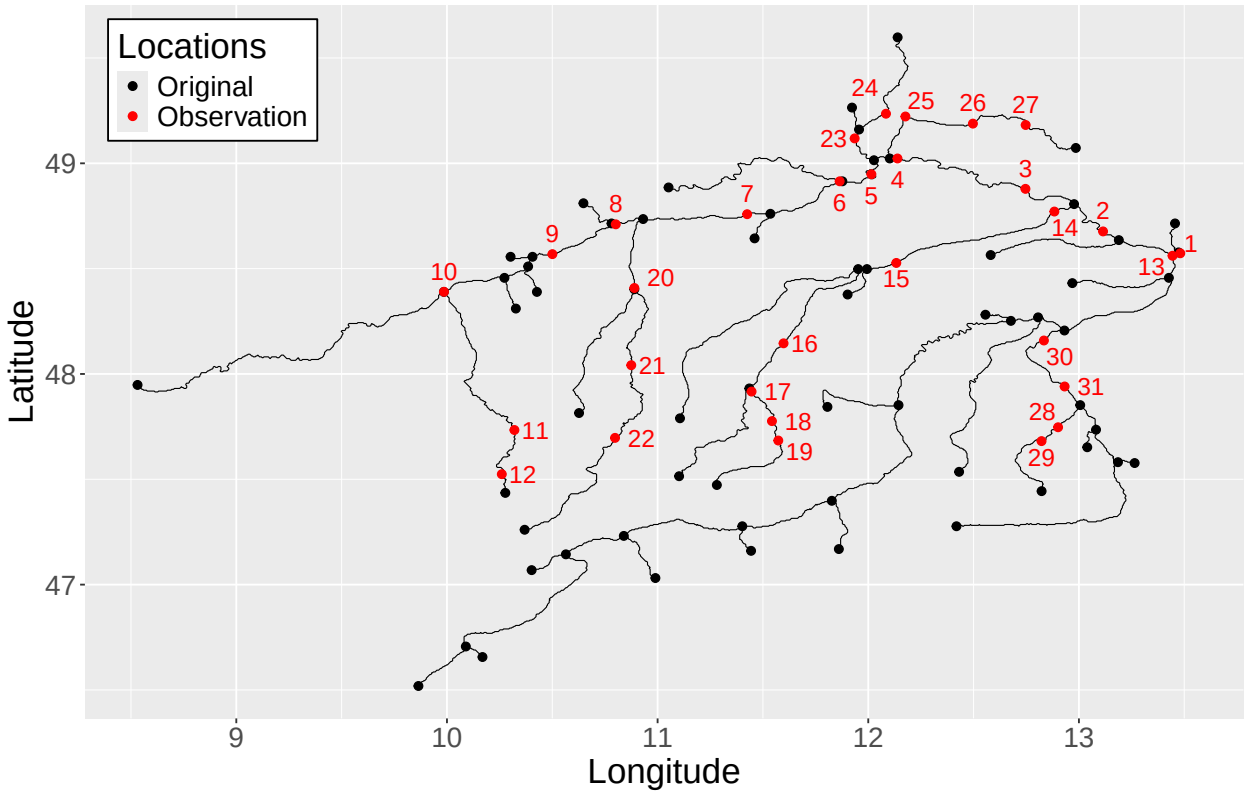

Figure 4.3: River outline of the upper Danube River basin with black and red points representing the nodes which determine the graphical structure of the physical network (original) and observation locations, respectively. The red numbers correspond to the station number.

the original and observed locations, respectively. Note that the original vertices correspond to the confluence and end-points of the underlying river network. Any vertices which belong to both sets are removed from $\boldsymbol{v}$. Figure 4.3 depicts the two sets of vertices for the upper Danube River basin. Now let $\boldsymbol{W}_{v,s} = (\boldsymbol{W}_v, \boldsymbol{W}_s)$ be the random vector containing the process at the original $\boldsymbol{W}_v = \{W_{s_0}(v_i) : i = 1, \dots, m\}$ and observation $\boldsymbol{W}_s = \{W_{s_0}(s_j) : j = 1, \dots, d\}$ locations. Since the finite-dimensional marginal distribution of a Gaussian process is a multivariate Gaussian distribution, it follows that the distribution of $\boldsymbol{W}_{v,s}$ is multivariate Gaussian. From the specific definition of the Gaussian Matérn-Whittle field in equation (4.3.5), the distribution has zero mean and precision matrix

$$Q = \begin{bmatrix} Q_{vv} & Q_{vs} \\ Q_{sv} & Q_{ss} \end{bmatrix},$$

where, for the case that $\delta = 1$, the elements are defined in Bolin et al. (2023a, Corollary 3)

as

$$Q_{ij} = 2\kappa\tau^2 \begin{cases} \sum_{e\in\bar{\mathcal{E}}_{v_i}} \left(\frac{1}{2} + \frac{\exp(-2\kappa l_e)}{1-\exp(-2\kappa l_e)}\right) \mathbb{I}(\bar{e} \neq \underline{e}) + \tanh(\kappa l_e/2)\mathbb{I}(\bar{e} = \underline{e}) & \text{if } i = j, \\ \sum_{e\in\bar{\mathcal{E}}_{v_i}\cap\bar{\mathcal{E}}_{v_j}} -\frac{\exp(-2\kappa l_e)}{1-\exp(-2\kappa l_e)} & \text{if } i \neq j, \end{cases}$$

such that $\bar{\mathcal{E}}_v$ denotes the set of edges incident to vertex v, l_e is the length of the edge e, and $\bar{e}$ and $\underline{e}$ denote the vertices that start and end edge e, respectively.

Before presenting the likelihood function, we make three observations. Firstly, since data is available only for $\boldsymbol{W}_s$, the likelihood requires the d-dimensional marginal distribution obtained by integrating over $\boldsymbol{W}_v$. Secondly, to adapt this method for the residual process of the CMEVM, it requires the additional constraint that $W_{s_0}(s_0) = 0$. This requires us to derive the relevant finite-dimensional conditional density. Obtaining the analytical form for this matrix when $\delta > 1$ is beyond the scope of our contribution. In what follows, we restrict our attention to the case $\delta = 1$. Lastly, we include a zero-mean Gaussian measurement error term with variance ε^2. This is assumed to be independently distributed over space. We include this term for reasons similar to those expressed by Simpson and Wadsworth (2021), namely that it improved the inference on the other parameters. The term will be ignored when making predictions (see Section 4.3.3.)

We can now write $\boldsymbol{W}_s = A\boldsymbol{W}_{v,s} + \varepsilon^2 I_{d\times d}$, where $A = [\mathbf{0}_{d\times m}, I_{d\times d}]$ maps $\boldsymbol{W}_{v,s}$ to $\boldsymbol{W}_s$, $\mathbf{0}_{d\times m}$ is the $(d\times m)$ zero matrix, and $I_{d\times d}$ is the $(d\times d)$ identity matrix. Since the linear transformation of a Gaussian random vector is also Gaussian, it follows that $\boldsymbol{W}_s \sim \text{MVN}_d(\mathbf{0}, \Gamma^{-1})$ with precision matrix $\Gamma = \varepsilon^{-2} I_{d\times d} - \varepsilon^{-4} A(Q + \varepsilon^{-2} AA^T)^{-1} A^T$ (see Bolin et al. (2023a) for a detailed derivation). Next, conditioning on $W_{s_0}(s_0) = 0$, we obtain that $\boldsymbol{W}_{s_0} \mid (W_{s_0}(s_0) = 0) \sim \text{MVN}_{d-1}\left(\mathbf{0}, \Gamma_0^{-1}\right)$, where $\boldsymbol{W}_{s_0}$ is the $(d-1)$-dimensional vector with the component corresponding to s_0 removed, and Γ_0 denotes Γ excluding the row and column corresponding to the conditioning location s_0. We fit this model separately to each conditioning site $s_0 \in \mathcal{S}$ by numerical optimisation of the log-likelihood function

$$l(\kappa, \tau, \varepsilon \mid \boldsymbol{w}_0) = -\frac{(d-1)n_{s_0}}{2}\log(2\pi) - \frac{n_{s_0}}{2}\log\left(\left|\Gamma_0^{-1}\right|\right) - \frac{1}{2}\sum_{t=1}^{n_{s_0}} \boldsymbol{w}_{0,t}^T \Gamma_0 \boldsymbol{w}_{0,t}.$$

For ease of notation we have dropped the subscript s_0, instead writing $\boldsymbol{w}_0 = (\boldsymbol{w}_{0,1}, \ldots, \boldsymbol{w}_{0,n_{s_0}})$ where $\boldsymbol{w}_{0,t}$ is the $(d-1)$-dimensional vector of standardised residuals associated with the tth exceedance of $Y(s_0)$ by $u_{s_0}^Y$, and n_{s_0} is the number of such exceedances.

4.3.3 Prediction

To make predictions across the network, we can simulate values at prediction vertices, which are added to an extended graph after parameter estimation. The reason that the prediction vertices can be added after fitting the model is that, by construction, the field is invariant to the addition or removal of nodes of degree two. Consequently, provided that (a) all vertices corresponding to prediction locations have degree two and (b) inclusion of these additional vertices does not increase the degree of the existing vertices, then there is no need to refit the model with the prediction vertices included. This is in contrast to competitor models based on the graph Laplacian (Borovitskiy et al., 2021), which require a new fit every time a new set of prediction locations is added, since changes to the underlying graph alter the graph Laplacian.

The prediction process starts with the definition of a fine mesh over the network. Denoting the mesh locations as $\boldsymbol{p} = (p_1, \ldots, p_k)$, a new extended metric graph is defined as $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ with vertex set $\mathcal{V}^* = (\boldsymbol{v}^*, s_0)$ where $\boldsymbol{v}^* = (\boldsymbol{p}, \boldsymbol{v}, \boldsymbol{s} \backslash s_0)$ (Bolin et al., 2023b). Under the Gaussian field assumption, the finite-dimensional distribution of $\boldsymbol{W}_{v^*,s_0} = (\boldsymbol{W}_{v^*}, W_{s_0}(s_0))$, where $\boldsymbol{W}_{v^*} = \{W_{s_0}(v_l^*) : l = 1, \ldots, k+m+d-1\}$, is multivariate Gaussian distribution with zero mean and block precision matrix Q_{v^*,s_0}. Recall that we want to simulate $\boldsymbol{W}_{v^*} | W_{s_0}(s_0) = 0$. Using standard properties of the multivariate Gaussian distribution, this conditional distribution is also multivariate Gaussian with zero mean and precision matrix $Q_{v^*|s_0} = Q_{v^*v^*} - Q_{v^*s_0} Q_{s_0 s_0}^{-1} Q_{s_0 v^*}$. Simulation from this distribution is straightforward.

The full prediction procedure is as follows. First simulate $Y(s_0) \mid Y(s_0) > q_{s_0}^Y$ from a standard exponential distribution. Use this value to evaluate the normalising functions $\hat{\alpha}_{s|s_0}^{GAM}(\boldsymbol{c}(s))$, $\hat{\beta}_{s|s_0}^{GAM}(\boldsymbol{c}(s))$, $\hat{\mu}_{s|s_0}^{GAM}(\boldsymbol{c}(s))$ and $\hat{\sigma}_{s|s_0}^{2GAM}(\boldsymbol{c}(s))$, manually enforcing these at the conditioning location, i.e. set $\hat{\alpha}_{s|s_0}^{GAM}(\boldsymbol{c}(s_0)) = 1$ and so forth. Combine these values with the simulation of $\boldsymbol{W}_{v^*,s_0}$ to obtain $\{Y_{s_0}(s)\} \coloneqq \{(Y(i), Y(s_0)) \mid Y(s_0) > q_{s_0}^Y : i \in \boldsymbol{v}^*\}$ by first reversing equation (4.3.4) and then reversing equation (4.3.3). Finally, we transform the process back onto the original margins using a double application of the PIT. We achieve this by transforming the process onto uniform margins, denoted $\{U_{s_0}(s)\}$, using the standard Laplace distribution before applying the inverse of the CDF of the non-stationary GEV in equation (4.3.1). The result is a simulated surface for the process $\{X_{s_0}(s)\} \coloneqq \{(X(i), X(s_0)) \mid X(s_0) > q_{s_0}^X : i \in \boldsymbol{v}^*\}$ after an exponential transformation. Note that $q_{s_0}^X$ is also transformed onto the original margins using the same procedure, and predictions can be made for any threshold $q_{s_0}^Y \geq u_{s_0}^Y$.

In the continuous spatial setting, Richards et al. (2022) use an importance sampling procedure to obtain predictions for the unconditioned process $\{X(s) : s \in \mathcal{S}\}$. The procedure requires sampling from the observed data when the maximum of the process at time t does not exceed

some threshold $u' \geq \max_{s_0 \in \mathcal{S}}\{u^X_{s_0}\}$. Such sampling is not possible for unobserved locations. While they suggest this can be overcome using infilling techniques, such as quantile regression (Fasiolo et al., 2021), such techniques are not required in our case since we can simulate from the non-stationary GEV in equation (4.3.1) with probability

$$\mathbb{P}\left[\max_{s\in\mathcal{S}}\{X(s)\} \leq u'\right] = \frac{1}{n}\sum_{t=1}^{n} \max\{\boldsymbol{x}_t\} \leq u'.$$

Sampling from the non-stationary GEV may require computationally expensive rejection sampling to ensure the maximum does not exceed u'. Given this, we do not obtain predictions for the unconditioned process $\{X(s) : s \in \mathcal{S}\}$, but note that it is possible.

4.4 Application

We apply the model described in Section 4.3 to the upper Danube River basin dataset introduced in Section 4.1. Recall that for this dataset, there are $d = 31$ observation locations and $n = 428$ independent spatial events. We detail the modelling steps, as well as model diagnostics, before providing predictions over the entire river network.

4.4.1 Marginal modelling

We first model the marginal distributions using the non-stationary GEV model in equation (4.3.1). The primary covariates are latitude and longitude, but we also construct more specific covariates to fully capture the trends induced by the river structure and local topography. All splines are univariate, except for those that include latitude and longitude, which use a 2-dimensional spline, to better capture changes in the parameters due to confluences in the river network. When fitting the splines, the choice of basis dimension $D_k \in \mathbb{N}$ is pivotal since it determines the degrees of freedom, or flexibility, of the spline. We follow the approach of taking D_k to be larger than expected and allow the smoothing parameters to control the smoothness (Youngman, 2022). For example, if the data suggested that a cubic spline could be appropriate, we would set D_k to approximately 6 to ensure a sufficient number of knots and that the spline is flexible enough to capture the non-linear trends in the data.

A forward selection process, shown in Table 4.1, determined the best model for the non-stationary GEV to include longitude and latitude in all parameters. In addition, the location and scale parameters include river distance from station 6 (see Figure 4.3), while the shape parameter has a separate intercept for the "Regen" tributary, which contains stations $25-27$. The spline for longitude and latitude implicitly captures the Euclidean distance and direction

Table 4.1: Table of selected models considered for the non-stationary GEV in equation (4.3.1). Indicator functions are denoted by $\mathbb{I}\{\cdot\}$, coefficients are denoted γ_i^* for $i \in \mathbb{N}_0$, thin-plate regression splines of dimension k with respect to covariate $c(s)$ at location $s \in \mathcal{S}$ are denoted $s_k^*(c_j(s))$, and $c_j(s)$ for $j \in \{1, 2, 3, 4\}$ correspond to the longitude, latitude, river distance from station 6, and tributary name, respectively, at location $s \in \mathcal{S}$. The change in AIC/BIC from model 1 has been provided to the nearest integer for each model. Bold values denote the model with the lowest AIC/BIC (excluding the saturated model (model 8)).

Model	$\upsilon(\boldsymbol{c}(s))$	$\psi(\boldsymbol{c}(s))$	$\xi(\boldsymbol{c}(s))$	Number of Parameters	Δ AIC	Δ BIC
1	γ_0^υ	γ_0^ψ	γ_0^ξ	3	0	0
2	$\gamma_0^\upsilon + s_{25}^\upsilon(c_1(s),\ c_2(s))$	γ_0^ψ	γ_0^ξ	27	$-21,509$	$-21,330$
3	$\gamma_0^\upsilon + s_{25}^\upsilon(c_1(s),\ c_2(s))$	$\gamma_0^\psi + s_{20}^\psi(c_1(s),\ c_2(s))$	γ_0^ξ	46	$-22,273$	$-21,955$
4	$\gamma_0^\upsilon + s_{25}^\upsilon(c_1(s),\ c_2(s))$	$\gamma_0^\psi + s_{20}^\psi(c_1(s),\ c_2(s))$	$\gamma_0^\xi + s_{15}^\xi(c_1(s),\ c_2(s))$	60	$-22,412$	$-21,999$
5	$\gamma_0^\upsilon + s_{25}^\upsilon(c_1(s),\ c_2(s))$	$\gamma_0^\psi + s_{20}^\psi(c_1(s),\ c_2(s)) + s_{15}^\psi(c_3(s))$	$\gamma_0^\xi + s_{15}^\xi(c_1(s),\ c_2(s))$	74	$-22,626$	$-22,153$
6	$\gamma_0^\upsilon + s_{25}^\upsilon(c_1(s),\ c_2(s)) + s_{15}^\upsilon(c_3(s))$	$\gamma_0^\psi + s_{20}^\psi(c_1(s),\ c_2(s)) + s_{15}^\psi(c_3(s))$	$\gamma_0^\xi + s_{15}^\xi(c_1(s),\ c_2(s))$	88	$-24,426$	$-23,940$
7	$\gamma_0^\upsilon + s_{25}^\upsilon(c_1(s),\ c_2(s)) + s_{15}^\upsilon(c_3(s))$	$\gamma_0^\psi + s_{20}^\psi(c_1(s),\ c_2(s)) + s_{15}^\psi(c_3(s))$	$\gamma_0^\xi + \gamma_1^\xi \mathbb{I}\{c_4(s) = \text{“Regen”}\} + s_{15}^\xi(c_1(s),\ c_2(s))$	89	$\mathbf{-24,446}$	$\mathbf{-23,959}$
8	$\gamma_0^\upsilon + \sum_{i=2}^{31} \gamma_{i-1}^\upsilon \mathbb{I}\{c_4(s) = i\}$	$\gamma_0^\psi + \sum_{i=2}^{31} \gamma_{i-1}^\psi \mathbb{I}\{c_4(s) = i\}$	$\gamma_0^\xi + \sum_{i=2}^{31} \gamma_{i-1}^\xi \mathbb{I}\{c_4(s) = i\}$	93	$-24,450$	$-23,776$

between stations and therefore performs better than a spline just based on Euclidean distance. River distance is clearly important (see Asadi et al. (2015) for more details). While we could use a different river distance for each conditioning site, we found that using an anchored location in the centre of the domain performed better on average. Finally, a separate intercept was included for the "Regen" tributary as it drastically improved the fit. This could be due to the previous splines not having enough knot locations in this region, so the behaviour of the "Naab" tributary (the tributary containing stations 23 - 24) is dominating the splines. Further, the geographical terrain and catchment areas of the sites north of the main Danube River are drastically different compared to stations north of the river, which could explain why a separate intercept is necessary. While an intercept for both northern tributaries was considered, we found that using a spline for one tributary was better.

Figure 4.4 shows the predicted parameter estimates from model 7 in Table 4.1 over the entire river network. The shape parameter is largely negative, except for the tributaries with stations $25 - 27$ and $17 - 19$. While most of the predicted parameter values lie within plausible ranges, there were a small number of unrealistically low/high values of the location/scale parameters. These correspond to the "Inn" tributary, which corresponds to the dark blue/grey region for the location/scale parameter in Figure 4.4. Based on the spatial location and the lack of data in this region, these unrealistic estimates are unsurprising and can be attributed to spline extrapolation. Splines become linear beyond the range of the observed data. For this region, the locations correspond to the lowest latitudes and those furthest away in terms of river distance from station 6. Thus, both splines are extrapolating, resulting in unrealistic estimates. That being said, this does not invalidate estimates for the rest of the region, as they closely align with the estimates from the stationary GPD at the site level.

The non-stationary GEV is highly parametrised, with 89 (64 effective) parameters compared

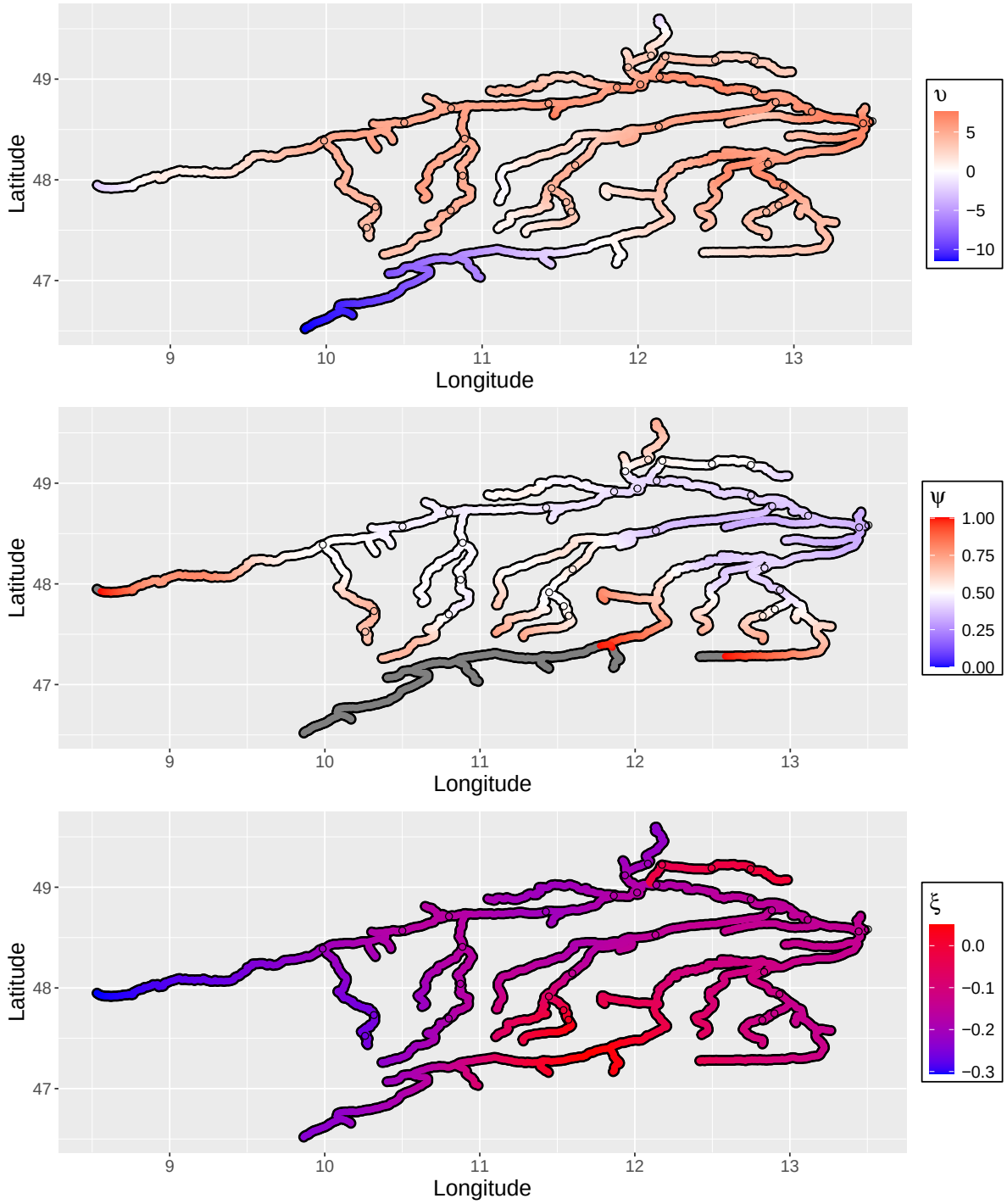

Figure 4.4: Spatial plots of predicted parameter estimates from model 7 in Table 4.1. The panels from top to bottom correspond to the location, scale, and shape parameters for the non-stationary GEV in equation (4.3.1). In each panel, the colour corresponds to the predicted parameter estimate at that location. The circles on the river network correspond to observation locations.

to 28 parameters in the regionalised model of Asadi et al. (2015). Given that there are 93 parameters in the saturated model (model 8 in Table 4.1), there is a strong possibility that the chosen model is over-fitting to the data. Evidence for this can be seen in the AIC and BIC: the BIC for model 7 is lower than that of the saturated model, and the AICs for the two models are almost identical. Further, the QQ-plots for the two models are similar for the majority of stations (see right panel of Figure 4.5). However, the fits of the two models do exhibit differences at some sites (see left and centre panels of Figure 4.5). For station 17, model 7 overestimates the empirical quantiles while model 8 better captures the tail at the expense of midrange quantiles. For station 18, model 7 overestimates the most extreme empirical quantiles, but model 8 explodes for this station. Since the discrepancy between models 7 and 8 is localised to a small section of the river, we overlooked the differences. However, the GEV, and thereby a GAM based on the GEV, may not be appropriate. This is because the multivariate identification process (see Section 4.5 for more details) does not necessarily provide declustered maxima at each location across the river network. Thus, the GAMS in this region may be overcompensating for the incorrect distributional assumption.

4.4.2 Dependence modelling

CMEVM Parameters

We first model the spatial dependence structure of the data using the original CMEVM model as per Keef et al. (2013), with the dependence threshold $u^Y_{s_0}$ set to the 0.8-quantile of the standard Laplace distribution. Next, we model the fitted parameter estimates $\hat{*}^{HT}_{s|s_0}$ (see Section 4.3.2) using GAMs. As discussed in Section 4.3.2, the GAMs are fitted separately for each conditioning station $s_0 \in \mathcal{S}$. To enable direct comparability between conditioning sites, we use the same covariates for all conditioning stations. A covariate is included only if it reduces the AIC for at least 50% of the conditioning locations. Table 4.2 details the model selection process for each of the CMVEM parameters.

Since a separate GAM is fitted to each of the conditioning stations and each GAM has $15-19$ parameters for 31 "observations", there is a risk of over-fitting. Reassuringly, Figure 4.6 shows that the GAM predictions are in overall good agreement with the pointwise estimates, but that there is still some noise in the GAM predictions. While the GAMs have a large number of parameters, there are fewer parameters than the pairwise method, and we can spatially interpolate the CMEVM parameters over the entire river network.

Figure 4.7 shows the predicted GAM estimates over the river network when we condition on station 28 experiencing a large event. The predictions for $\alpha_{s|s_0}(\boldsymbol{c}(s))$ and $\beta_{s|s_0}(\boldsymbol{c}(s))$ have very similar patterns, and as expected the former is close to 1 and the latter is approximately 0 for

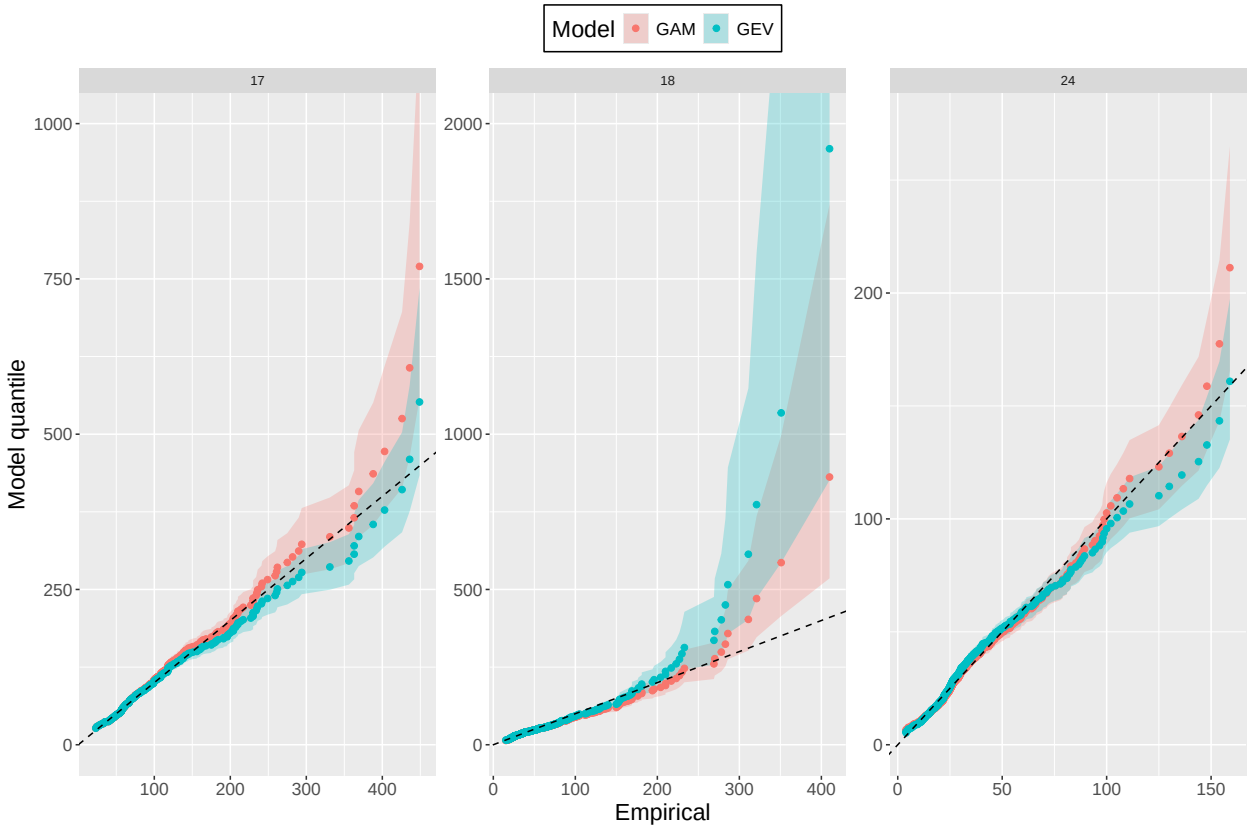

Figure 4.5: QQ-plots for selected stations comparing the empirical and model quantiles from models 8 (GEV) and 7 (GAM) in Table 4.1 represented in red and blue, respectively. The shaded regions correspond to 95% tolerance bounds for the model quantiles from 500 non-parametric bootstrapped samples. The model points correspond to the median across said samples. The black dashed lines represent the $y = x$ line.

river locations near the conditioning site. Consequently, the predicted values for $\mu_{s|s_0}(\boldsymbol{c}(s))$ are close to 0 for river locations near the conditioning site. However, $\sigma^2_{s|s_0}(\boldsymbol{c}(s))$ is closer to 0.5, rather than the expected value of 1. This is because we have not imposed constraints on the spline, for example, that $\sigma^2_{s|s_0}(\boldsymbol{c}(s))$ should be 1 at the conditioning location. Thus, the GAM is smoothing out the artificial "observation" of $\sigma^{2HT}_{s_0|s_0}(\boldsymbol{c}(s)) = 1$ to fit to the surrounding data points. In any case, the predicted values for each parameter largely decay smoothly with distance from the conditioning location, except for the "Naab" and "Regen" tributaries, which contain stations $23 - 24$ and $25 - 27$, respectively, where the separate intercepts cause a discontinuity.

Conditioning on the other stations yields similar findings except for some oddities on the most southern tributary (the Inn), which is likely caused by the tributary having no data, being far from the observation locations and being at a much higher altitude than much of the network. Otherwise, the predictions are as expected.

Table 4.2: GAMs model selection for CMEVM parameters. Notation follows from Table 4.1 except $c_j(s)$ for $j \in \{1,2,3,4,5\}$ are covariates corresponding to (from the conditioning location) river distance, longitude difference, latitude difference, Euclidean distance, and tributary name, respectively, for $s \in \mathcal{S}$. AIC/BIC Lower details the number of times the AIC/BIC is lower compared to the previous model. RMSE details the root mean squared error (1dp) in the $\hat{*}^{HT}_{s|s_0}$ and $\hat{*}^{GAM}_{s|s_0}(\boldsymbol{c}(s))$ over all d models.

Model	$\alpha_{s\|s_0}(\boldsymbol{c}(s))$	Number of Parameters	AIC Lower	BIC Lower	RMSE
1	γ_0^α	1	–	–	5.9
2	$\gamma_0^\alpha + s_5^\alpha(c_1(s))$	5	31	31	4.0
3	$\gamma_0^\alpha + s_5^\alpha(c_1(s)) + s_5^\alpha(c_3(s))$	9	24	19	3.1
4	$\gamma_0^\alpha + s_5^\alpha(c_1(s)) + s_5^\alpha(c_2(s)) + s_5^\alpha(c_3(s))$	13	29	18	2.6
5	$\gamma_0^\alpha + \gamma_1^\alpha\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\alpha(c_1(s)) + s_5^\alpha(c_2(s)) + s_5^\alpha(c_3(s))$	14	17	13	2.2
6	$\gamma_0^\alpha + \gamma_1^\alpha\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\alpha(c_1(s)) + s_5^\alpha(c_2(s)) + s_5^\alpha(c_3(s)) + s_5^\alpha(c_4(s))$	18	19	17	2.2
7	$\gamma_0^\alpha + \gamma_1^\alpha\mathbb{I}\{c_5(s) = \text{“Naab”}\} + \gamma_2^\alpha\mathbb{I}\{c_5(s) = \text{“Regen”}\} + s_5^\alpha(c_1(s)) + s_5^\alpha(c_2(s)) + s_5^\alpha(c_3(s)) + s_5^\alpha(c_4(s))$	19	18	18	2.0
Model	$\beta_{s\|s_0}(\boldsymbol{c}(s))$	Number of Parameters	AIC Lower	BIC Lower	RMSE
1	γ_0^β	1	–	–	4.2
2	$\gamma_0^\beta + \gamma_1^\beta\mathbb{I}\{c_5(s) = \text{“Naab”}\}$	2	20	16	3.8
3	$\gamma_0^\beta + \gamma_1^\beta\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\beta(c_3(s))$	6	25	19	3.2
4	$\gamma_0^\beta + \gamma_1^\beta\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\beta(c_2(s)) + s_5^\beta(c_3(s))$	10	22	15	2.8
5	$\gamma_0^\beta + \gamma_1^\beta\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\beta(c_1(s)) + s_5^\beta(c_2(s)) + s_5^\beta(c_3(s))$	14	23	15	2.4
6	$\gamma_0^\beta + \gamma_1^\beta\mathbb{I}\{c_5(s) = \text{“Naab”}\} + \gamma_2^\beta\mathbb{I}\{c_5(s) = \text{“Regen”}\} + s_5^\beta(c_1(s)) + s_5^\beta(c_2(s)) + s_5^\beta(c_3(s))$	15	18	17	2.4
Model	$\mu_{s\|s_0}(\boldsymbol{c}(s))$	Number of Parameters	AIC Lower	BIC Lower	RMSE
1	γ_0^μ	1	–	–	6.1
2	$\gamma_0^\mu + s_5^\mu(c_1(s))$	5	31	24	4.8
3	$\gamma_0^\mu + s_5^\mu(c_1(s)) + s_5^\mu(c_3(s))$	9	24	22	4.0
4	$\gamma_0^\mu + s_5^\mu(c_1(s)) + s_5^\mu(c_2(s)) + s_5^\mu(c_3(s))$	13	26	16	3.2
5	$\gamma_0^\mu + \gamma_1^\mu\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\mu(c_1(s)) + s_5^\mu(c_2(s)) + s_5^\mu(c_3(s))$	14	25	20	2.8
6	$\gamma_0^\mu + \gamma_1^\mu\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\mu(c_1(s)) + s_5^\mu(c_2(s)) + s_5^\mu(c_3(s)) + s_5^\mu(c_4(s))$	18	19	15	2.5
Model	$\sigma^2_{s\|s_0}(\boldsymbol{c}(s))$	Number of Parameters	AIC Lower	BIC Lower	RMSE
1	γ_0^σ	1	–	–	9.1
2	$\gamma_0^\sigma + s_5^\sigma(c_3(s))$	5	30	26	7.0
3	$\gamma_0^\sigma + s_5^\sigma(c_1(s)) + s_5^\sigma(c_3(s))$	9	26	18	6.3
4	$\gamma_0^\sigma + \gamma_1^\sigma\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\sigma(c_1(s)) + s_5^\sigma(c_3(s))$	10	26	22	5.2
5	$\gamma_0^\sigma + \gamma_1^\sigma\mathbb{I}\{c_5(s) = \text{“Naab”}\} + s_5^\sigma(c_1(s)) + s_5^\sigma(c_3(s)) + s_5^\sigma(c_4(s))$	14	21	16	4.7
6	$\gamma_0^\sigma + \gamma_1^\sigma\mathbb{I}\{c_5(s) = \text{“Naab”}\} + \gamma_2^\sigma\mathbb{I}\{c_5(s) = \text{“Regen”}\} + s_5^\sigma(c_1(s)) + s_5^\sigma(c_3(s)) + s_5^\sigma(c_4(s))$	15	20	18	4.3
7	$\gamma_0^\sigma + \gamma_1^\sigma\mathbb{I}\{c_5(s) = \text{“Naab”}\} + \gamma_2^\sigma\mathbb{I}\{c_5(s) = \text{“Regen”}\} + s_5^\sigma(c_1(s)) + s_5^\sigma(c_2(s)) + s_5^\sigma(c_3(s)) + s_5^\sigma(c_4(s))$	19	18	14	4.0

Residual process

After fitting the GAMs to the CMEVM parameters, we can transform the data to obtain the standardised residual process in equation (4.3.4). As discussed in Section 4.3.2, we model the process using the Gaussian Whittle-Matérn field on a metric graph with $\delta = 1$ under both the direct and indirect observation methods. Since we model the process $\{W_{s_0}(s)\}$ for d conditioning locations, we will obtain d MLEs. However, the MLEs are not very informative, and it is better to assess the model fit directly.

Figure 4.8 compares the empirical $(d-1) \times (d-1)$ conditional correlation matrix for $\{W_{s_0}(s) : s \in \boldsymbol{s} \backslash s_0\}$ and the corresponding matrix for the direct and indirect models. When we condition on station 28 (left panel), the direct model consistently underestimates the empirical correlations, while the indirect model captures the correlation matrix well. The same comparison is made when $s_0 = 4$ (right panel). Neither model captures the empirical correlation matrix well in this case for several reasons. First, the assumption that the standardised residual process is a zero-mean Gaussian process is less suitable for $s_0 = 4$ compared to $s_0 = 28$, which is a knock-on effect from assuming symmetry for the margins of the resid-

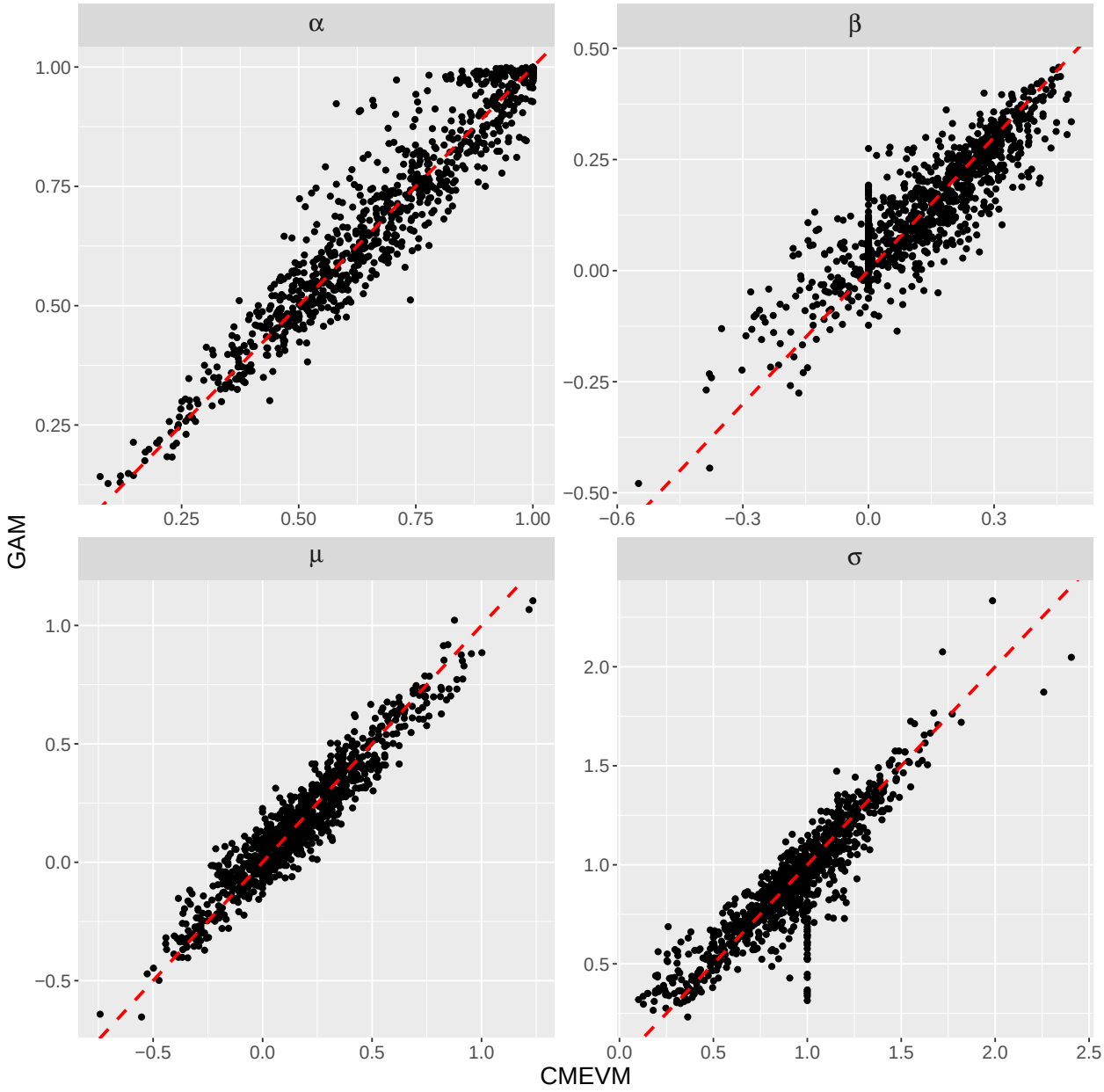

Figure 4.6: Scatter plots comparing parameter estimates ($\alpha_{s|s_0}$ (top left), $\beta_{s|s_0}$ (top right), $\mu_{s|s_0}$ (bottom left), and $\sigma^2_{s|s_0}$ (top right)) from the original CMEVM and estimates from the GAMs in Table 4.2. The red dashed lines represent the $y = x$ line.

ual process. Second, and more pertinently, the model performs poorly for some conditioning locations due to the simplifying assumption that $\delta = 1$ in equation (4.3.5). When $\delta = 1$, the model is unable to capture negative conditional correlations (see Supplementary Material), which are present when conditioning on some gauging stations, e.g., station 4.

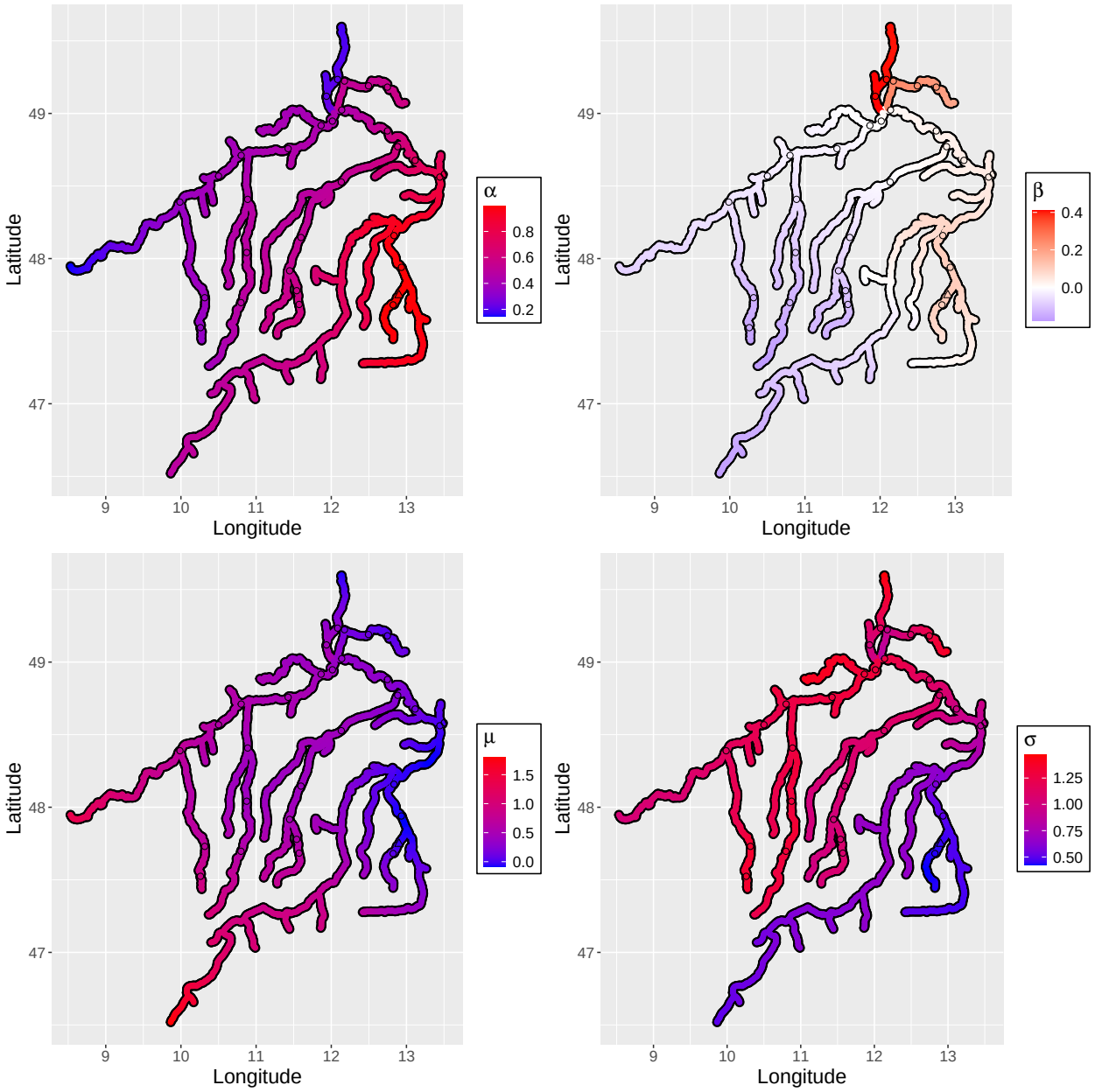

Figure 4.7: Spatial plots of predicted CMEVM parameters using the GAMs in Table 4.2. The panels correspond to the $\alpha_{s|s_0}(\boldsymbol{c}(s))$ (top left), $\beta_{s|s_0}(\boldsymbol{c}(s))$ (top right), $\mu_{s|s_0}(\boldsymbol{c}(s))$ (bottom left), and $\sigma^2_{s|s_0}(\boldsymbol{c}(s))$ (bottom right) when $s_0 = 28$ (the triangle on the network). The colour corresponds to the predicted parameter estimate at that location. The circles on the river network correspond to the other observation locations.

4.4.3 Prediction

Despite the residual process providing a poor fit for some of the conditioning stations, we will perform prediction and assess the predictive qualities of the model. First, we will assess spatial surfaces for the entire river network, which, due to the step-wise inference procedure,

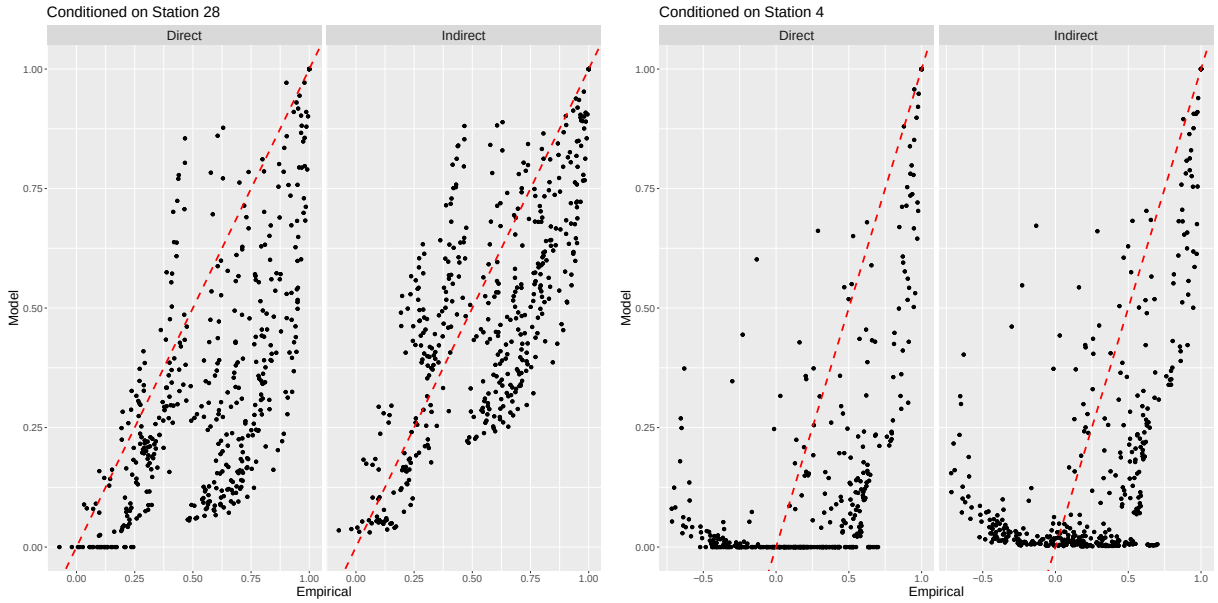

Figure 4.8: Scatter plots comparing the empirical conditional correlation matrix and the corresponding matrix from the direct and indirect models, when we condition on stations 28 (left) and 4 (right). The red dashed lines represent the $y = x$ line.

can be analysed for each stage of the model $\{W_{s_0}(s)\}$, $\{Z_{s_0}(s)\}$, $\{Y_{s_0}(s)\} \coloneqq \{Y(s) \mid Y(s_0) > u^Y_{s_0} : s \in \mathcal{S}\}$, and $\{\log(X_{s_0}(s))\} \coloneqq \{\log(X(s)) \mid \log(X(s_0)) > \log(u^X_{s_0}) : s \in \mathcal{S}\}$ to disentangle the contributions from the different model components. Such surfaces allow us to visualise the spatial dependence in river flows over the network. However, they do not quantify the predictive qualities of our model. We will assess this using summary statistics and pairwise sample clouds.

Spatial surfaces

Figure 4.9 shows the pointwise mean (left column) and standard error (right column) for 1000 simulations from the fitted indirect model when we condition on station 28 (see Section 4.3.3 for details). Assessing the standard error first, it is encouraging to see that this is relatively small for all stages of the model. In addition, the standard errors for locations close to the conditioning site are close to 0. Naturally, the standard error generally increases smoothly as we go further away from the conditioning station. The exception to this is for $\{\log(X_{s_0}(s))\}$ where the most southern tributary (the Inn) has a large standard error, despite being close to the conditioning location. This is caused by the large estimates of $\psi(\boldsymbol{c}(s))$ for this region.

While we would expect $\{W_{s_0}(s)\}$ to change smoothly with river distance, the top left panel of Figure 4.9 ungulates because we are analysing the pointwise mean of the simulated surfaces

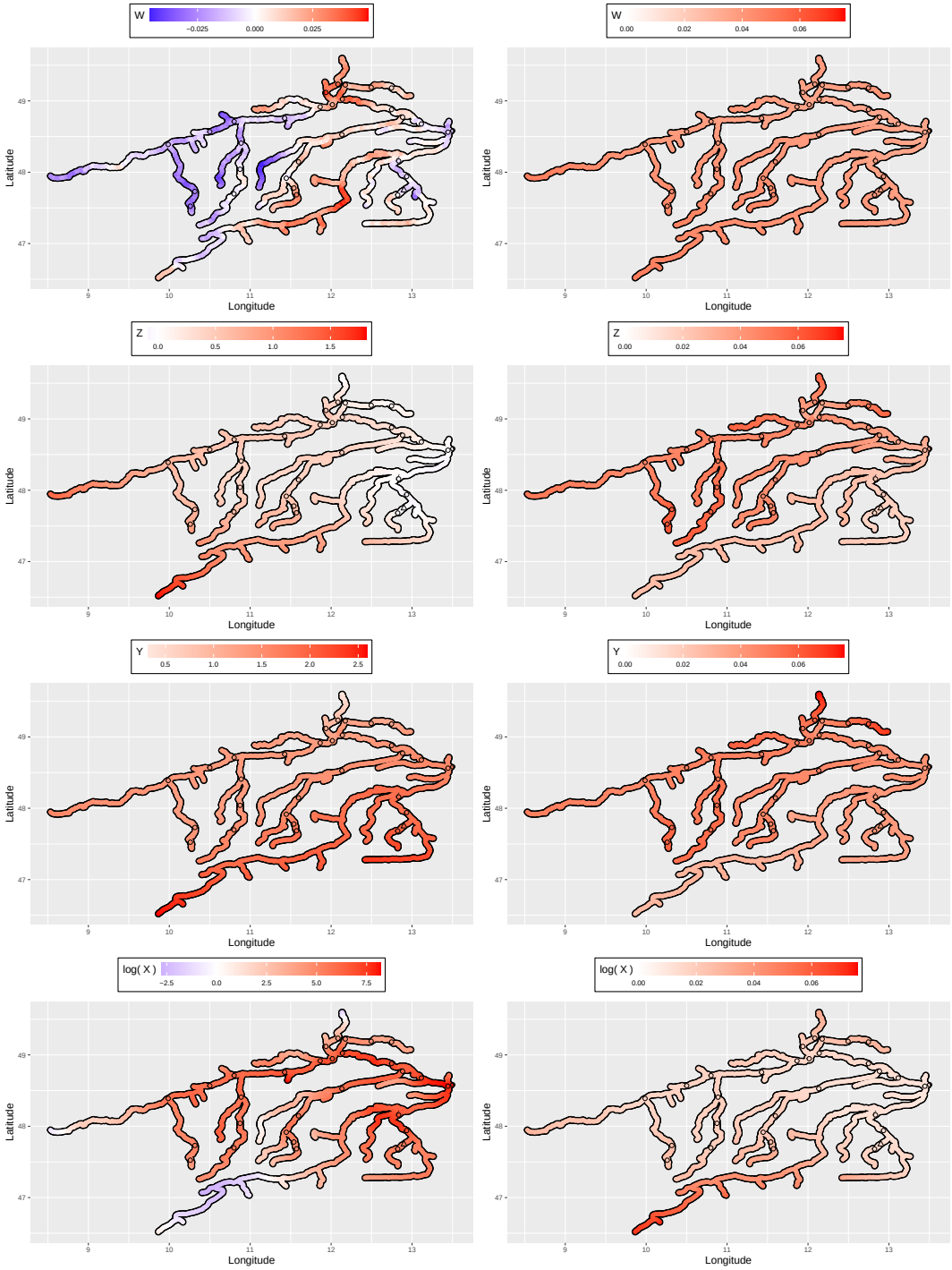

Figure 4.9: Spatial plots of mean (left) and standard error (right) over 1000 simulated river flow surfaces from the fitted indirect model when we condition on station 28 (the triangle on the network) being large. This is shown for each step in the modelling procedure, $\{W_{s_0}(s)\}$, $\{Z_{s_0}(s)\}$, $\{Y_{s_0}(s)\}$, and $\{\log(X_{s_0}(s))\}$, from top to bottom. In each panel, the colour corresponds to the mean/standard error of the river flow at that location. The circles on the river network correspond to the other observation locations.

rather than a single simulated surface. This could be addressed by assessing simultaneous confidence bands rather than pointwise estimates, but we don't do so here. The main thing to note is that the average value is tightly centred around 0. For $\{Z_{s_0}(s)\}$, the mean is approximately 0 around the conditioning location (a consequence of $\alpha_{s|s_0}(\boldsymbol{c}(s))$ and $\beta_{s|s_0}(\boldsymbol{c}(s))$ being close to 1 and 0, respectively, in this area (see Figure 4.7)). Since $\sigma^2_{s|s_0}(\boldsymbol{c}(s))$ is relatively small for all $s \in \mathcal{S}$ when we condition on station 28, the mean of $\{Z_{s_0}(s)\}$ is effectively dominated by $\mu_{s|s_0}(\boldsymbol{c}(s))$, resulting in a very similar spatial pattern to the bottom left panel of Figure 4.7. Moving onto $\{Y_{s_0}(s)\}$, again since $\beta_{s|s_0}(\boldsymbol{c}(s))$ is approximately 0 for majority of $s \in \mathcal{S}$, the spatial pattern looks very similar to that of $\alpha_{s|s_0}(\boldsymbol{c}(s))$ (top left panel of Figure 4.7) except for the "Naab" and "Regen" tributaries where the mean is larger due to the high $\beta_{s|s_0}(\boldsymbol{c}(s))$ values in this region. Finally, for $\{\log(X_{s_0}(s))\}$, the mean on the Inn is slightly negative due to the negative values of $\upsilon(\boldsymbol{c}(s))$ on this tributary. Otherwise, we see that conditional on an extreme event at station 28, the mean flow around the conditioning location and towards the sink is relatively high. Furthermore, the main Danube River also has a relatively high mean, while the mean is relatively lower on the northern tributaries.

Summary statistics

While the maps in Figure 4.9 allow us to visualise the spatial dependence of river flows over the upper Danube River basin, they do not assess the quality of the predictions from the model. Figure 4.10 compares empirical and model-based estimates of the coefficient of tail dependence (Coles et al., 1999) $\{\eta_q(s_1, s_2) : s_1, s_2 \in \boldsymbol{s}; q \in \{0.8, 0.85, 0.9\}\}$ for $\{X(s) \mid X(s_0) > u^X_{s_0} : s, s_0 \in \boldsymbol{s}\}$ when s_0 is station 4 (top row) and station 28 (bottom row). To obtain a like-for-like comparison between the data and the simulations from the model, the model-based estimates are based on simulations using the same number of events used to fit the CMEVM dependence model. Although it is difficult to assess, the estimates are more centred around the $y = x$ line when we condition on station 4 than when we condition on station 28, despite the residual process providing a poorer fit for the former. In any case, the model tends to underestimate the empirical values, suggesting it is not simulating enough joint extreme events. However, there is significant variability in the estimates, which is partly due to these estimates being drawn from a single small (less than 100) sample in both cases.

Sample clouds

Given the variability in the estimates of η, we compare sample clouds of simulations from the fitted model and the data used to fit the model. Figure 4.11 shows these clouds for stations 14 and 19, and stations 11 and 27, when we condition on stations 4 and 28 being large. For

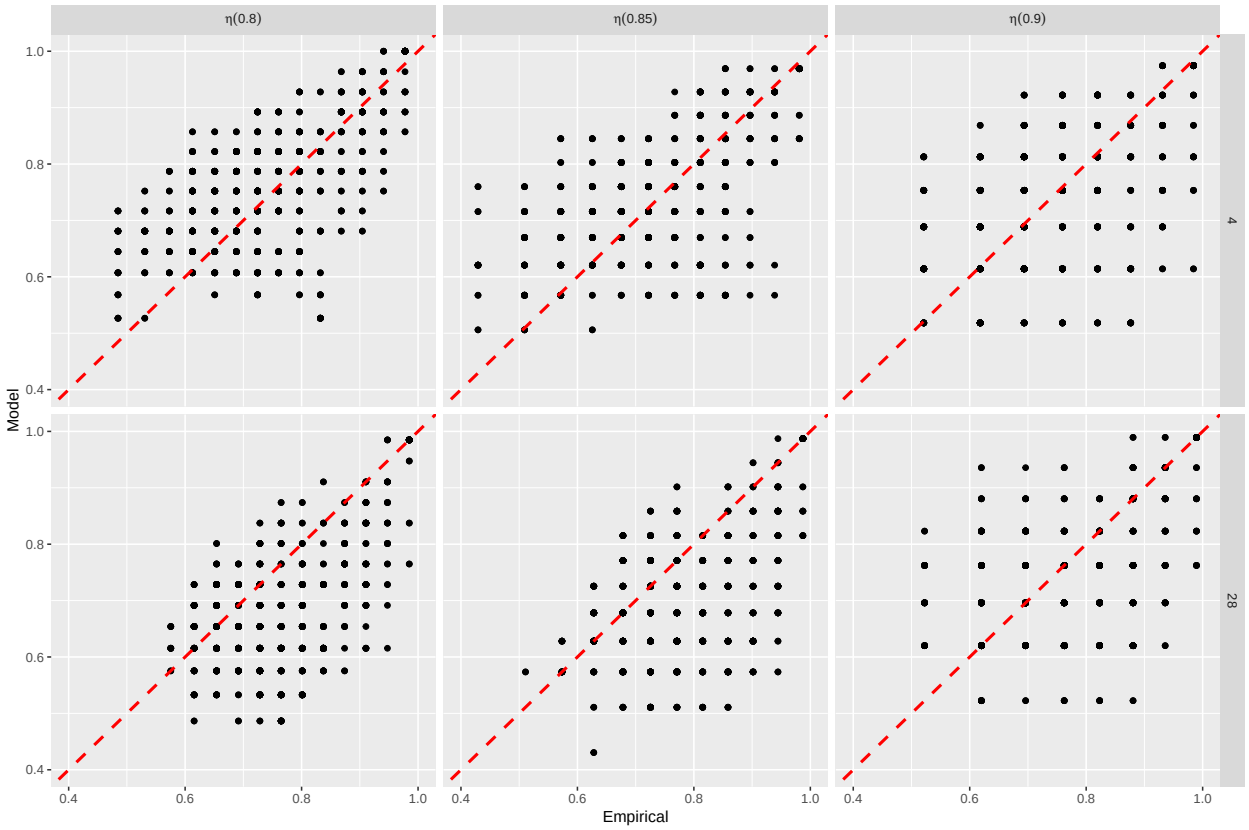

Figure 4.10: Empirical and model-based estimates of $\{\eta_q(s_1, s_2) : s_1, s_2 \in \boldsymbol{s}\}$ for $\{X(s) \mid X(s_0) > u^X_{s_0} : s, s_0 \in \boldsymbol{s}\}$ such that $q \in \{0.8, 0.85, 0.9\}$ (left to right). The top and bottom rows correspond to $s_0 = 4$ and $s_0 = 28$, respectively. The red dashed lines represent the $y = x$ line.

stations 14 and 19 (first two columns), the simulated river flows do not follow the trend of the data when we condition on station 4. This stems from the original data having a strong positive correlation, while the model suggests the two stations are almost uncorrelated. While we observe some improvement when we condition on station 28, we notice the simulated river flows overestimate the dependence in the left tail and underestimate it in the right tail. For stations 11 and 27 (last two columns), the standardised residuals have negative (positive) correlation when we condition on station 4 (28). Despite this, the simulations from the model appear somewhat reasonable in the top three rows. However, in the last two rows, the model is simulating too many samples when at least one component is not large (best seen on uniform margins in the fourth row).

For all the examples provided, our inclination is that the discrepancies between the simulated and observed data can be largely attributed to: i) the simplifying assumption that the margins of the residual process are symmetric, ii) the simplifying assumption that $\delta = 1$ in the Gaussian-Whittle Matérn process. For i), kernel density estimators of the marginal

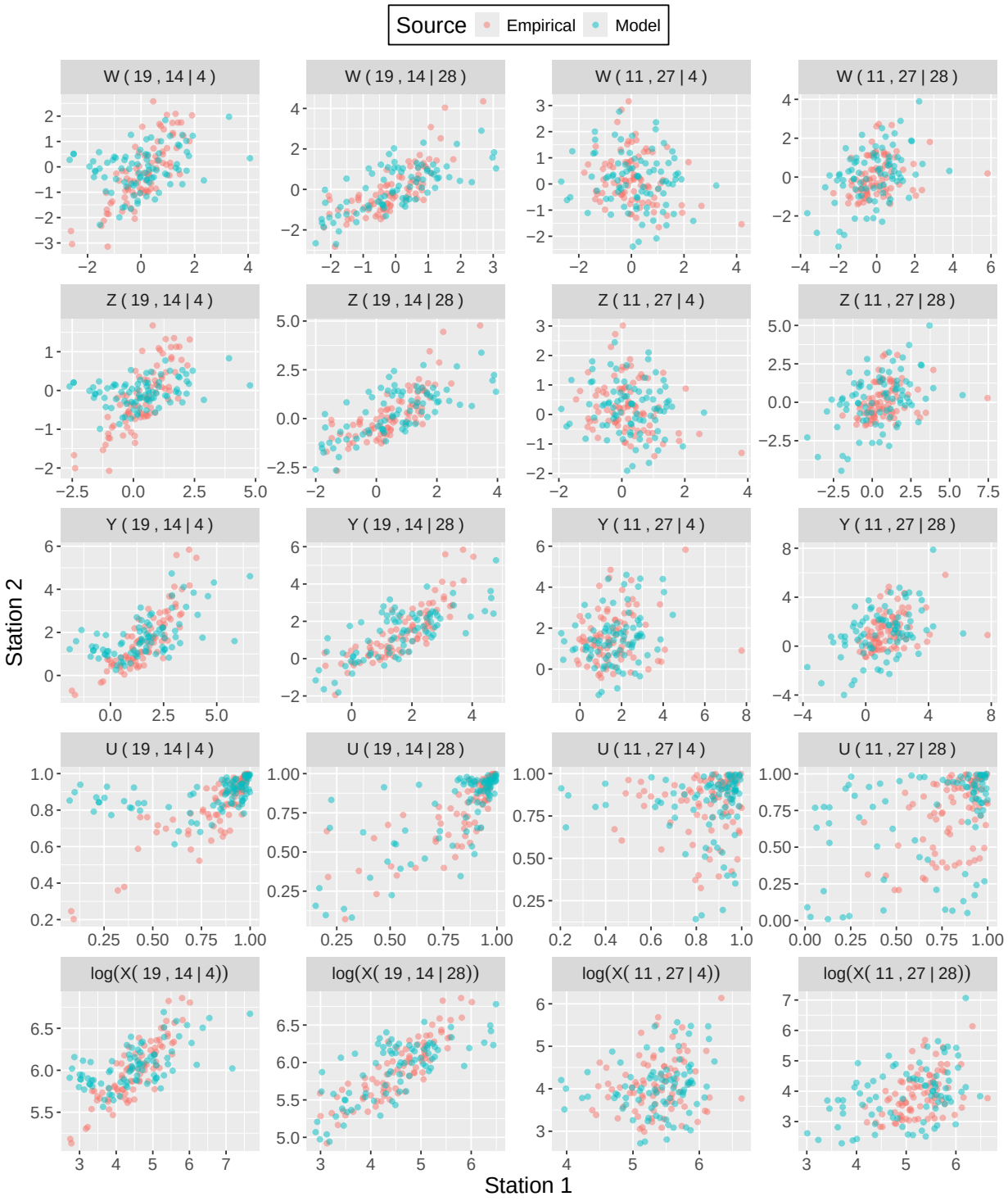

Figure 4.11: Selected bivariate sample clouds comparing a single simulated sample (blue) from the fitted indirect model and the data used to fit the model (orange). The title format is $*(s_1, s_2 \mid s_0)$, where $*$ corresponds to the margins, s_1 is plotted on the $x-$axis, s_2 is plotted on the $y-$axis, and s_0 is the conditioning station. The margins of the process are $\{W_{s_0}(s)\}$, $\{Z_{s_0}(s)\}$, $\{Y_{s_0}(s)\}$, $\{U_{s_0}(s)\}$, $\{\log(X_{s_0}(s))\}$ from top to bottom.

distributions exhibit clear asymmetry. Ignoring this means our assumption that $\{W_{s_0}(s)\}$ is a zero-mean Gaussian process is inappropriate, which will impact the model fit and subsequent predictions. For ii), the assumption that the sample path is continuous but not differentiable substantially restricts the correlation structure, which limits the range of simulations from the model.

4.5 Discussion

This contribution is the first attempt, to our knowledge, to obtain spatial predictions over an entire river network using an extreme value model that can capture both AD and AI. While the results in Section 4.4 are preliminary, we believe that the methodology has potential. In this section, we outline what we have learnt so far about the proposed method, its limitations, and how it might be improved.

A fundamental challenge with the CMEVM is the requirement to condition on $W_{s_0}(s_0) = 0$. Direct conditioning requires a closed-form solution for the precision matrix of the Gaussian Whittle-Matérn field on a metric graph, which we could obtain only for $\delta = 1$. Consequently, the Matérn correlation function in equation (4.1.1) degenerates to the exponential correlation function, which meant we were unable to capture the empirical negative conditional correlations in the process $W_{s_0}(s) \mid \big(W_{s_0}(s_0) = 0\big)$. This resulted in poor predictive performance when we conditioned on some stations. Extending the closed form for the precision matrix when $\delta = 2$ would extend the scope of the model to capture negative conditional correlations, improving the model fit (Figure 4.8), and improving the prediction accuracy from the model (Figure 4.11). However, it may be possible to circumvent the conditioning. For instance, Wadsworth and Tawn (2022) suggest the constraint could be imposed by modelling the process $\{W_{s_0}(s) : s \in \mathcal{S}\}$ and subtracting $W_{s_0}(s_0)$. Alternatively, Simpson et al. (2023) suggest conditioning by kriging (Cressie, 1993) could be adopted by generating a realisation of the conditional distribution subject to the constraint $W_{s_0}(s_0) = 0$. Either of these methods could avoid explicit conditioning and, in turn, eliminate the need for a closed-form solution for the precision matrix when $\delta \in \mathbb{N}$. Therefore, we recommend investigating both to assess the feasibility of their implementation, and their impact on the model fit if they can be implemented.

Conditioning may be avoided altogether by using alternative extreme value models, such as random scale-mixture models (Huser et al., 2017) or models based on geometric extremes (Nolde and Wadsworth, 2022). For example, Hazra et al. (2025) propose the sparse Gaussian spatial-scale mixture process such that $X(s) = R(s)Z(s)$ for $s \in \mathcal{S} \subset \mathbb{R}^2$, where $R(s)$ is

the random scale component that is positive with probability 1 for all $s \in \mathcal{S}$, and $Z(s)$ is a standard Gaussian process. The random scale component is extremely flexible and controls the level of extremal dependence over the spatial domain. While they model $Z(s)$ in a continuous spatial setting using a GMRF on a finite mesh, we hypothesise that $Z(s)$ could be modelled using a GMRF on a metric graph to model processes on non-Euclidean spaces. Alternatively, models based on geometric extremes have recently been extended to the continuous spatial setting in unpublished work. Again, we hypothesise that this could be extended to processes on non-Euclidean spaces by replacing the Gaussian process component with a GMRF on a metric graph. If the above approaches to circumvent the conditioning in the CMEVM cannot be implemented, then we suspect these models could be utilised for modelling joint extreme events on a network.

Returning to our model, there are several areas for improvement. For example, we could allow for asymmetry in the margins of the residual process (Farrell et al., 2024) to ensure the standardised residual process is closer to a standard white noise process. This would improve the fit of the GMRF on the metric graph component and the subsequent simulations from the model. Another extension would be to include the river structure in the estimation of the parameters. For instance, exploratory data analysis of $\hat{\alpha}^{HT}_{s|s_0}$ found that the parameters tended to decay as one descends/ascends a tributary. Such behaviour could be captured by using a combination of the tail-up and tail-down models proposed by Ver Hoef and Peterson (2010), a variance component model (Asadi et al., 2015; Ver Hoef and Peterson, 2010), or a latent Gaussian modelling framework (Jóhannesson et al., 2022) with GMRFs on metric graphs (Bolin et al., 2024).

Another issue with our model is the potential over-fitting. For the margins, the fitted GEV GAM model has 89 (64 effective) parameters, which is close to the number of parameters (93) in the saturated model. The parameter space could be reduced by strategically placing the knots at confluence points in the river (e.g. at the black point between stations 2 and 14 in Figure 4.3) since we expect a change here due to the change in the volume of river flow. Adaptive splines could be used to learn the knot locations, potentially reducing the parameter space, but they are slow to fit. Alternatively, we could use a regionalised model. These models, which account for hydrological similarity and spatial proximity (Fischer and Schumann, 2021), are standard practice in many hydrological applications. Asadi et al. (2015) used such an approach in their marginal model, which only required 28 marginal parameters. However, such models require the catchment area to be known for all observation and prediction locations. Deriving this is beyond the scope of our contribution, but it could be investigated in the future. There is also a danger of the GAM models over-fitting the CMEVM parameters. To mitigate against this we could investigate alternative metrics to

add covariates to the GAMs, for instance combined k-fold cross-validation (Hastie et al., 2001, Chapter 7) with the continuous ranked probability score (CRPS) (Gneiting and Katzfuss, 2014) to minimise the difference in the observed values and the predicted distribution function without the need to specify empirical quantiles (André et al., 2025).

The two-step fitting process of the CMEVM parameters could be avoided by utilising `EVGAM` (Youngman, 2022). However, the `EVGAM` model often failed to estimate $\beta_{s|s_0}(\boldsymbol{c}(s))$ well when it was fitted to each conditioning site. One possible way to improve estimation is to assume spatial stationarity, which would allow pooling of information between the conditioning sites (Wadsworth and Tawn, 2022). Given the clear non-stationarity in our case (Figure 4.2), this was not explored further. We suspect the $\beta_{s|s_0}(\boldsymbol{c}(s))$ is poorly estimated either because the covariates are uninformative or a consequence of the joint estimation of the CMEVM parameters. For the former, utilising covariates based on hydrological catchment characteristics may be useful. For the latter, inference could be stabilised by adopting a step-wise procedure in which $\beta_{s|s_0}(\boldsymbol{c}(s))$ is first set to 0 and $\alpha_{s|s_0}(\boldsymbol{c}(s))$, $\mu_{s|s_0}(\boldsymbol{c}(s))$, and $\sigma^2_{s|s_0}(\boldsymbol{c}(s))$ are estimated. Next, $\alpha_{s|s_0}(\boldsymbol{c}(s))$ is fixed to the estimate in the previous step, and $\beta_{s|s_0}(\boldsymbol{c}(s))$, $\mu_{s|s_0}(\boldsymbol{c}(s))$, and $\sigma^2_{s|s_0}(\boldsymbol{c}(s))$ are estimated. The idea is to ensure $\alpha_{s|s_0}(\boldsymbol{c}(s))$ corresponds to the line of best fit while $\beta_{s|s_0}(\boldsymbol{c}(s))$ captures the variability in the data. We have implemented this procedure in the multivariate setting with promising results, but have yet to extend it to the spatial setting.

Although the non-stationary GEV is a suitable model for the majority of stations, it is not appropriate for stations 17 - 19 (Figure 4.5). Since these stations are confined to a single small tributary of the river network, we overlooked this. The poor marginal fits may be an artefact of the event identification process; Asadi et al. (2015) decluster the data over the entire spatial region using a nine-day moving window until no windows remain. This declustering scheme has several undesirable consequences. Firstly, because the declustering is performed over the entire river network, it can result in the removal of some events that would have been detected if the declustering had been performed at the station level, which can potentially adversely affect the marginal models. Secondly, since there is no lower bound on the flow that constitutes a joint extreme event, the flow at some stations is not necessarily extreme for all events. Consequently, a threshold-based marginal model may be more appropriate than a GEV. Thirdly, the declustered data results in a dependence structure that is closer to full asymptotic dependence than may be suggested by the raw data. This is in part due to the global declustering and the conservative nine-day window choice. Given that the declustering scheme directly impacts both the marginal behaviour and the dependence structure, in addition to investigating alternative marginal models, the sensitivity to both a lower event bound and window length should be investigated.

We found (not shown) that threshold-based extreme value models provided good marginal fits at all stations. Obtaining the threshold over the entire spatial domain could be achieved by either 1) using a GAM to model the pointwise thresholds (a two-step approach), or 2) using quantile regression (Fasiolo et al., 2021). However, to obtain uncensored predictions of river flow over the entire network, we would also require a model below the threshold. While mixture models exist within the extremes literature (Scarrott and MacDonald, 2012), we believe these have yet to be applied to the non-stationary setting. Given the extension is not trivial, a simpler approach would be to use a non-stationary GEV with an appropriate event identification process at the station level described above.

In conclusion, we propose that a full spatio-temporal model be explored for the raw data, such that the temporal dependence accounts for the lag in extreme events due to the time it takes for water to travel downstream on the river. Spatio-temporal models for river networks have been considered in the geostatistics literature by O'Donnell et al. (2014) to model pollution levels and by Santos-Fernandez et al. (2022) to model temperature. The former employs flexible splines for modelling seasonality and spatial and temporal dependence, whereas the latter utilises a Bayesian hierarchical model that accounts for dependence through vector autoregressive models. Both models are fitted to all the data, meaning they are unlikely to capture the dependence of the extremes well. However, the ideas could be incorporated into extreme value models. For instance, Simpson et al. (2023) use a latent Gaussian modelling framework to extend the CMEVM to the spatio-temporal setting for modelling the surface temperature of the Red Sea. To extend this to non-Euclidean spaces, the spatial dependence could include two GMRFs, one based on Euclidean/hydrological distance and one based on river distance. Although implementing a spatio-temporal model for data collected on a network may be challenging, we believe this is the most promising option for accurately modelling and predicting extreme river flow events.

Supplementary Material to "Conditional Extremes with Metric Graphs"

S4.1 Direct observations method

For direct observations (in which there is no measurement error), the random vector containing the process at the observation locations, denoted $\boldsymbol{W}_s = \{W(s_j) : j = 1, \ldots, d\}$, follows a multivariate Gaussian distribution. Specifically, $\boldsymbol{W}_s \sim \text{MVN}_d(\boldsymbol{0}, Q_s^{-1})$, such that

$$Q_s = Q_{ss} - Q_{sv}Q_{vv}^{-1}Q_{vs} = \begin{bmatrix} Q_{s_*s_*} & Q_{s_*s_0} \\ Q_{s_0s_*} & Q_{s_0s_0} \end{bmatrix},$$

where $s_* = \boldsymbol{s} \backslash s_0$ (Bolin et al., 2023a, Corollary 5). Conditioning on $W_{s_0}(s_0) = 0$, it follows that $\boldsymbol{W}_{s_0} \mid (W_{s_0}(s_0) = 0) \sim \text{MVN}_{d-1}\left(\boldsymbol{0}, Q_{s_*}^{-1}\right)$, where $\boldsymbol{W}_{s_0}$ is the $(d-1)$-dimensional vector with the component corresponding to s_0 removed, and $Q_{s_*} = Q_{s_*s_*} - Q_{s_*s_0}Q_{s_0s_0}^{-1}Q_{s_0s_*}$.

S4.2 Conditional correlations

When $\delta = 1$, the correlation function for a Gaussian Whittle-Matérn field on a metric graph, detailed in equation (4.1.1), reduces to the exponential correlation function. Consequently, the model is unable to obtain negative conditional correlations. To see this, assume the random vector $\boldsymbol{X} = (X_1, X_2, X_3)$ follows a zero-mean Gaussian process with a correlation matrix that can be represented by the graphical structure $1 - 2 - 3$. Let $\rho_{ij} \in [-1, 1]$ and $d_{ij} \geq 0$ represents the correlation between components X_i and X_j and the distance between locations i and j, respectively, for $i, j \in \{1, 2, 3\}$. For metric graphs, the distance is defined as the shortest path along the graph, meaning $d_{13} = d_{12} + d_{23}$.

Conditioning on the first component, the conditional covariance matrix is

$$\text{Cov}(X_2, X_3 \mid X_1) = \begin{pmatrix} 1 - \rho_{12}^2 & \rho_{23} - \rho_{12}\rho_{13} \\ \rho_{23} - \rho_{12}\rho_{13} & 1 - \rho_{13}^2 \end{pmatrix}.$$

Thus, for some rate parameter $\lambda > 0$

$$\begin{aligned} \rho_{23} - \rho_{12}\rho_{13} &= \exp(-\lambda d_{23}) - \exp(-\lambda d_{12})\exp(-\lambda d_{13}) \\ &= \exp(-\lambda d_{23}) - \exp(-\lambda d_{12})\exp(-\lambda(d_{12} + d_{23})) \\ &= \exp(-\lambda d_{23})(1 - \exp(-2\lambda d_{12})) \geq 0, \end{aligned}$$

since distance is non-negative. Similar results can be obtained when we condition on the third component due to symmetry.

Conditioning on the second component, the conditional covariance matrix is

$$\mathrm{Cov}(X_1, X_3 \mid X_2) = \begin{pmatrix} 1-\rho_{13}^2 & \rho_{13}-\rho_{12}\rho_{23} \\ \rho_{13}-\rho_{12}\rho_{23} & 1-\rho_{23}^2 \end{pmatrix}.$$

Again, for some rate parameter $\lambda > 0$

$$\begin{aligned} \rho_{13}-\rho_{12}\rho_{23} &= \exp(-\lambda d_{13}) - \exp(-\lambda d_{12})\exp(-\lambda d_{23}) \\ &= \exp(-\lambda(d_{12}+d_{23})) - \exp(-\lambda d_{12})\exp(-\lambda d_{23}) \\ &= \exp(-\lambda d_{12})\exp(-\lambda d_{23}) - \exp(-\lambda d_{12})\exp(-\lambda d_{23}) \\ &= 0. \end{aligned}$$

Thus, components X_1 and X_3 are conditionally independent given X_2.

These results apply to larger trees, as we will always be conditioning on an end node or a node between two other nodes in the graph. Therefore, we are unable to obtain negative conditional correlations in the $\delta = 1$ model.

Bibliography

André, L. M., Campbell, R., D'Arcy, E., Farrell, A., Healy, D., Kakampakou, L., Murphy, C., Murphy-Barltrop, C. J. R., and Speers, M. (2025). Extreme value methods for estimating rare events in Utopia. *Extremes*, 28(1):23–45.

Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023 – 2050.

Bergström, S. (1991). Principles and confidence in hydrological modelling. *Hydrology Research*, 22(2):123–136.

Bolin, D., Simas, A., and Wallin, J. (2023a). Statistical inference for Gaussian Whittle-Matérn fields on metric graphs. *arXiv preprint arXiv:2304.10372*.

Bolin, D., Simas, A. B., and Wallin, J. (2023b). *MetricGraph: Random fields on metric graphs*. R package version 1.4.0.

Bolin, D., Simas, A. B., and Wallin, J. (2024). Gaussian Whittle–Matérn fields on metric graphs. *Bernoulli*, 30(2):1611 – 1639.

Borovitskiy, V., Azangulov, I., Terenin, A., Mostowsky, P., Deisenroth, M., and Durrande, N. (2021). Matérn Gaussian Processes on Graphs. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2593–2601. PMLR.

Brunner, M. I., Furrer, R., and Favre, A.-C. (2019). Modeling the spatial dependence of floods using the Fisher copula. *Hydrology and Earth System Sciences*, 23(1):107–124.

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(2):127.

Cressie, N. A. C. (1993). *Statistics for spatial data*. John Wiley & Sons, Inc., New York, revised edition.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical Modeling of Spatial Extremes. *Statistical Science*, 27(2):161 – 186.

de Fondeville, R. and Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika*, 105(3):575–592.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer, New York.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932.

Engelke, S., Hitz, A. S., Gnecco, N., and Hentschel, M. (2025). *graphicalExtremes: Statistical Methodology for Graphical Extreme Value Models*. R package version 0.3.4.

Farrell, A., Eastoe, E. F., and Lee, C. (2024). Conditional extremes with graphical models. *arXiv preprint arXiv:2411.17013*.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. (2021). Fast Calibrated Additive Quantile Regression. *Journal of the American Statistical Association*, 116(535):1402–1412.

Fischer, S. and Schumann, A. H. (2021). Regionalisation of flood frequencies based on flood type-specific mixture distributions. *Journal of Hydrology X*, 13:100107.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2001). *The elements of statistical learning : data mining, inference, and prediction*. Springer Series in Statistics. Springer, New York, NY, 1 edition.

Hazra, A., Huser, R., and Bolin, D. (2025). Efficient modeling of spatial extremes over large geographical domains. *Journal of Computational and Graphical Statistics*, 34(3):795–811.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Huser, R., Opitz, T., and Thibaud, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatial Statistics*, 21:166–186.

Huser, R. and Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(1):e1537.

Jóhannesson, Á. V., Siegert, S., Huser, R., Bakka, H., and Hrafnkelsson, B. (2022). Approximate Bayesian inference for analysis of spatiotemporal flood frequency data. *The Annals of Applied Statistics*, 16(2).

Keef, C., Papastathopoulos, I., and Tawn, J. A. (2013). Estimation of the conditional dis-

tribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, 115:396–404.

Keef, C., Tawn, J., and Svensson, C. (2009). Spatial Risk Assessment for Extreme River Flows. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 58(5):601–618.

Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., Dunning, P., and Batstone, C. (2010). A new method to assess the risk of local and widespread flooding on rivers and coasts. *Journal of Flood Risk Management*, 3(4):323–336.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Luppichini, M., Vailati, G., Fontana, L., and Bini, M. (2024). Machine learning models for river flow forecasting in small catchments. *Scientific Reports*, 14(1):26740.

Merz, R. and Blöschl, G. (2005). Flood frequency regionalisation — spatial proximity vs. catchment attributes. *Journal of Hydrology*, 302(1):283–306.

Monte, L. D., Huser, R., Papastathopoulos, I., and Richards, J. (2025). Generative modelling of multivariate geometric extremes using normalising flows. *arXiv preprint arXiv:2505.02957*.

Murphy-Barltrop, C. J. R., Majumder, R., and Richards, J. (2024). Deep learning of multivariate extremes via a geometric representation. *arXiv preprint arXiv:2406.19936*.

Nolde, N. and Wadsworth, J. L. (2022). Linking representations for multivariate extremes via a limit set. *Advances in Applied Probability*, 54(3):688–717.

O'Donnell, D., Rushworth, A., Bowman, A. W., Scott, E. M., and Hallard, M. (2014). Flexible regression models over river networks. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 63(1):47–63.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards, J., Tawn, J. A., and Brown, S. (2022). Modelling extremes of spatial aggregates of precipitation using conditional methods. *The Annals of Applied Statistics*, 16(4):2693 – 2713.

Rohrbeck, C. and Cooley, D. (2023). Simulating flood event sets using extremal principal components. *The Annals of Applied Statistics*, 17(2):1333 – 1352.

Rohrbeck, C. and Tawn, J. A. (2021). Bayesian spatial clustering of extremal behavior for hydrological variables. *Journal of Computational and Graphical Statistics*, 30(1):91–105.

Ross, E., Sam, S., Randell, D., Feld, G., and Jonathan, P. (2018). Estimating surge in extreme North Sea storms. *Ocean Engineering*, 154:430–444.

Santos-Fernandez, E., Ver Hoef, J. M., Peterson, E. E., McGree, J., Isaak, D. J., and Mengersen, K. (2022). Bayesian spatio-temporal models for stream networks. *Computational Statistics & Data Analysis*, 170:107446.

Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *REVSTAT-Statistical Journal*, 10(1):33–60.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1):33–44.

Shooter, R., Tawn, J., Ross, E., and Jonathan, P. (2021). Basin-wide spatial conditional extremes for severe ocean storms. *Extremes*, 24(2):241–265.

Simpson, E. S., Opitz, T., and Wadsworth, J. L. (2023). High-dimensional modeling of spatial and spatio-temporal conditional extremes using INLA and Gaussian Markov random fields. *Extremes*, 26(4):669–713.

Simpson, E. S. and Wadsworth, J. L. (2021). Conditional modelling of spatio-temporal extremes for Red Sea surface temperatures. *Spatial Statistics*, 41:100482.

Towe, R. P., Tawn, J. A., Lamb, R., and Sherlock, C. G. (2019). Model-based inference of conditional extreme value distributions with hydrological applications. *Environmetrics*, 30(8):e2575.

Ver Hoef, J. M., Peterson, E., and Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological statistics*, 13(4):449–464.

Ver Hoef, J. M. and Peterson, E. E. (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, 105(489):6–18.

Wadsworth, J. and Tawn, J. (2022). Higher-dimensional spatial extremes via single-site conditioning. *Spatial Statistics*, 51:100677.

Wheater, H. S., Chandler, R. E., Onof, C., Isham, V. S., Bellone, E., Yang, C., Lekkas, D.,

Lourmas, G., and Segond, M.-L. (2005). Spatial-temporal rainfall modelling for flood risk estimation. *Stochastic Environmental Research and Risk Assessment*, 19(6):403–416.

Wood, S. N. (2017). *Generalized Additive Models*. Chapman and Hall/CRC.

Youngman, B. (2025). *evgam: Generalised Additive Extreme Value Models*. R package version 1.0.1, commit 9a87396d0e78921b61391655d1d5ad42f7727eef.

Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software*, 103(3):1–26.

Chapter 5

EVA (2023) Conference Data Challenge

Extreme value methods for estimating rare events in Utopia

EVA (2023) Conference Data Challenge: Team Lancopula Utopiversity

Abstract

To capture the extremal behaviour of complex environmental phenomena in practice, flexible techniques for modelling tail behaviour are required. In this paper, we introduce a variety of such methods, which were used by the Lancopula Utopiversity team to tackle the EVA (2023) Conference Data Challenge. This data challenge was split into four challenges, labelled C1-C4. Challenges C1 and C2 comprise univariate problems, where the goal is to estimate extreme quantiles for a non-stationary time series exhibiting several complex features. For these, we propose a flexible modelling technique, based on generalised additive models, with diagnostics indicating generally good performance for the observed data. Challenges C3 and C4 concern multivariate problems where the focus is on estimating joint probabilities. For challenge C3, we propose an extension of available models in the multivariate literature and use this framework to estimate joint probabilities in the presence of non-stationary dependence. Finally, for challenge C4, which concerns a 50-dimensional random vector, we employ a clustering technique to achieve dimension reduction and use a conditional modelling approach to estimate extremal probabilities across independent groups of variables.

Keywords: Extremal Dependence, Generalised Additive Modelling, Non-stationary Extremes, Peaks-over-threshold Modelling

5.1 Introduction

This paper details an approach to the data challenge organised for the Extreme Value Analysis (EVA) 2023 Conference. The objective of the challenge was to estimate extremal probabilities, or their associated quantiles, for simulated environmental data sets for various locations in a fictitious country called Utopia. The data challenge is split into 4 challenges; challenges C1 and C2 focus on a setting where data is obtained from a single location while challenges C3 and C4 concern multivariate data sets, where data is obtained simultaneously from multiple locations.

Challenge C1 requires estimation of the 0.9999-quantile of the distribution of the environmental response variable Y conditional on a covariate vector $\boldsymbol{X}$, for 100 realisations of covariates. To do so, we model the tail of $Y \mid \boldsymbol{X} = \boldsymbol{x}$ using a generalised Pareto distribution (GPD;

Pickands III, 1975) and employ the extreme value generalised additive modelling (EVGAM) framework, first introduced by Youngman (2019), to account for the non-stationary data structure. We consider a variety of model formulations and select our final model using cross-validation based on the continuous ranked probability score of the predicted quantiles. Furthermore, central 50% confidence intervals are estimated via a non-stationary bootstrapping technique, and the final model performance is assessed using the number of times the true conditional quantile lies in the confidence intervals (Rohrbeck et al., 2023). For Challenge C2, we are interested in estimating the value of q that satisfies $\Pr(Y > q) = 1/(300T)$, where $T = 200$.

Challenges C3 and C4 concern the estimation of probabilities for extreme multivariate regions, subsets of $\mathbb{R}^d$, where some or all of the components are so large that we seldom observe any data in them. Such estimates require techniques for modelling and extrapolating within the joint tail. For challenge C3, we want to estimate two joint tail probabilities for three unknown non-stationary environmental variables. To achieve this, we propose a non-stationary extension of the model introduced by Wadsworth and Tawn (2013). Lastly, for challenge C4, we wish to estimate the probability that 50 variables (locations) jointly exceed prespecified extreme thresholds. Based on an initial analysis, we separate the variables into five independent groups, and obtain distinct probability estimates for each group using the conditional extremes approach of Heffernan and Tawn (2004).

The remainder of the paper is structured as follows. A suitable background to EVA is provided in Section 5.2, introducing concepts required throughout our work. Section 5.3 covers our approach to the univariate challenges C1 and C2, and the multivariate challenges C3 and C4 are considered in Sections 5.4 and 5.5, respectively. The paper ends with a discussion of the results of all challenges in Section 5.6.

5.2 EVA background

5.2.1 Univariate modelling

Univariate EVA methods are concerned with capturing the behaviour of the tail of a distribution which allows for extreme quantities to be estimated. A common univariate approach is the peaks-over-threshold framework. Consider a continuous, independent and identically distributed (IID) random variable Y with distribution function F and upper endpoint $y^F := \sup\{y : F(y) < 1\}$. Pickands III (1975) shows that, for some high threshold $v < y^F$, the excesses $(Y - v) \mid Y > v$, after suitable rescaling, converge in distribution to a GPD as $v \to y^F$. Davison and Smith (1990) provide an overview of the properties of the GPD, and also

propose an extension of this framework to the non-stationary setting: given a non-stationary process Y with associated covariate(s) $\boldsymbol{X}$, the authors propose the following model

$$\Pr(Y > y + v \mid Y > v, \boldsymbol{X} = \boldsymbol{x}) = \left(1 + \xi(\boldsymbol{x})\frac{y}{\sigma(\boldsymbol{x})}\right)_{+}^{-1/\xi(\boldsymbol{x})}, \tag{5.2.1}$$

for $y > 0$, where $\sigma(\cdot), \xi(\cdot)$ are the covariate-dependent scale and shape parameters, respectively. Recent extensions of the Davison and Smith (1990) framework include allowing the threshold to be covariate-dependent, i.e., $v(\boldsymbol{x})$ (Kyselý et al., 2010; Northrop and Jonathan, 2011), and using generalised additive models (GAMs; Chavez-Demoulin and Davison, 2005; Youngman, 2019) to capture the functions $\sigma(\cdot)$ and $\xi(\cdot)$ in a flexible manner.

5.2.2 Extremal dependence measures

In addition to analysing marginal tail behaviours, multivariate EVA methods are concerned with quantifying the dependence between extremes of the individual components. An important classification of this dependence is obtained through the measure χ (Joe, 1997): given a d-dimensional random vector $\boldsymbol{Z}$, with $d \geq 2$ and $Z_i \sim F$ for all $i \in \{1, \ldots, d\}$,

$$\chi(u) := \left(\frac{1}{1-u}\right)\Pr(F(Z_1) > u, \ldots, F(Z_d) > u), \tag{5.2.2}$$

with $u \in [0, 1)$. Where the limit exists, we set $\chi := \lim_{u \to 1} \chi(u) \in [0, 1]$. When $\chi > 0$, we say that the variables in $\boldsymbol{Z}$ exhibit asymptotic dependence, i.e., can take their largest values simultaneously, with the strength of dependence increasing as χ approaches 1. If $\chi = 0$, the variables cannot all take their largest values together. In particular, for $d = 2$, we refer to the case $\chi = 0$ as asymptotic independence.

We also consider the coefficient of tail dependence proposed by Ledford and Tawn (1996). Using the formulation given in Resnick (2002), let

$$\eta(u) := \frac{\log(1-u)}{\log \Pr\left(F(Z_1) > u, \ldots, F(Z_d) > u\right)},$$

with $u \in [0, 1)$. When the limit exists, we set $\eta := \lim_{u \to 1} \eta(u) \in (0, 1]$. The cases $\eta = 1$ and $\eta < 1$, correspond to $\chi > 0$ and $\chi = 0$, respectively. For $\eta < 1$, this coefficient quantifies the form of dependence for random vectors that do not take their largest values simultaneously.

Since χ and η are limiting values, they are unknown in practice and must be approximated using numerical techniques. Therefore, when quantifying extremal dependence, we approximate $\chi(\eta)$ using empirical estimates of $\chi(u)(\eta(u))$ for some high threshold u.

5.3 Challenges C1 and C2

Both challenges concern 70 years of daily data for the capital city of Amaurot. Each year has 12 months of 25 days and two seasons (season 1 for months 1-6, and season 2 for months 7-12). Suppose Y is an unknown response variable, and $\boldsymbol{X} = (V_1, \ldots, V_8)$ is a vector of covariates, (V_1, V_2, V_3, V_4) denoting unknown environmental variables and (V_5, V_6, V_7, V_8) denoting season, wind direction (radians), wind speed (unknown scale), and atmosphere (recorded monthly), respectively.

For C1, we build a model for $Y \mid \boldsymbol{X}$ and estimate the 0.9999-quantile, with associated 50% confidence intervals, for 100 different covariate combinations denoted $\tilde{\boldsymbol{x}}_i$ for $i \in \{1, \ldots, 100\}$. Note $\tilde{\boldsymbol{x}}_i$ are not covariates observed within the data set, but new observations provided by the challenge organisers.

For C2, we estimate the marginal quantile q such that $\Pr(Y > q) = (6 \times 10)^{-4}$, which corresponds to a once in 200-year event in the IID setting; in particular, q is obtained subject to a predefined loss function. We first estimate the marginal distribution $F_Y(y)$ using Monte-Carlo techniques; see for instance, Eastoe and Tawn (2009). Since we have a large sample size, $n = 21,000$, it is reasonable to assume that the observed covariate sample is representative of $\boldsymbol{X}$. Thus, we can approximate the marginal distribution $F_Y(y)$ as follows,

$$\hat{F}_Y(y) = \int_{\boldsymbol{X}} F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{t=1}^{n} F_{Y_t|\boldsymbol{X}_t}(y_t \mid \boldsymbol{x}_t). \tag{5.3.1}$$

where $F_{Y|\boldsymbol{X}}(\cdot)$ is the conditional distribution function of $Y \mid \boldsymbol{X}$ and $f_{\boldsymbol{X}}(\cdot)$ denotes the joint probability density of the covariates $\boldsymbol{X}$.

We incorporate the following loss function provided by the challenge organisers,

$$\mathcal{L}(q, \hat{q}; \hat{\boldsymbol{\theta}}) = \begin{cases} 0.9(0.99q - \hat{q}) & \text{if } 0.99q > \hat{q}, \\ 0 & \text{if } |q - \hat{q}| \leq 0.01q, \\ 0.1(\hat{q} - 1.01q) & \text{if } 1.01q < \hat{q}, \end{cases} \tag{5.3.2}$$

where q and $\hat{q}$ are the true and estimated marginal quantiles, respectively. The estimated marginal quantiles $\hat{q}$ depend on the fitted marginal model parameters $\hat{\boldsymbol{\theta}}$. This loss function penalises under-estimation more heavily than an over-estimation.

We conduct the same exploratory data analysis for both challenges given the same covariates are used; this is outlined in Section 5.3.1. In Section 5.3.2 we introduce our techniques for modelling $Y \mid \boldsymbol{X}$, which is then used for modelling Y via (5.3.1). Our approach for uncertainty

quantification is outlined in Section 5.3.3, and we give our results for both challenges in Section 5.3.4.

5.3.1 Exploratory data analysis

The challenge editorial explains that the environmental response variable Y_t, $t \in \{1, \ldots, n\}$ is temporally independent, given the covariate vector $\boldsymbol{X}_t = \{V_{1,t}, \ldots, V_{8,t}\}$ (Rohrbeck et al., 2023), i.e., the response can be treated as independent, given some covaraites. However, it is not clear which covariates affect Y, and what form these covariate-response relationships take. In what follows, we aim to explore these relationships so we can account for them in our modelling framework.

To begin, we explore the dependence between all variables to understand the relationships between covariates, as well as the relationships between individual covariates and the response variable. We investigate dependence in the main body of the data using Kendall's τ measure, while for the joint tails, we use the pairwise extremal dependence coefficients χ and η defined in Section 5.2; values for all pairs are shown in Figure 5.1, with the threshold u set at the empirical 0.95-quantile for the extremal measures.

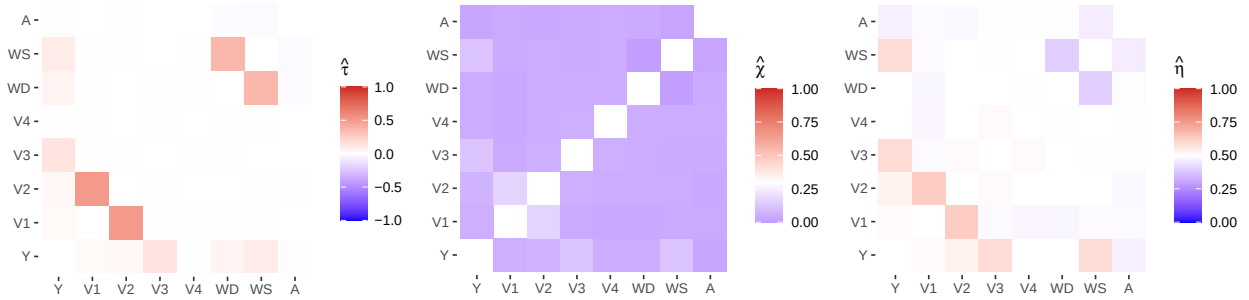

Figure 5.1: Heat maps for dependence measures for each pair of variables: Kendall's τ (left), χ (middle) and η (right). Note the scale in each plot varies, depending on the support of the measure, and the diagonals are left blank, where each variable is compared against itself.

The response variable Y has the strongest dependence with V_3 in the body of the distribution (see $\hat{\tau}$ in Figure 5.1), followed by V_6 (wind speed) then V_7 (wind direction). For the tail of the distribution, Y has strongest dependence with V_2, V_3 and V_6 (see $\hat{\chi}$ and $\hat{\eta}$ in Figure 5.1). We also find strong dependence between V_6 and V_7 in the body, but evidence of weak dependence in the tail (dark blue for $\hat{\chi}$ and $\hat{\eta}$). There is also strong dependence between V_1 and V_2 in both the body and tail (see dark red for $\hat{\eta}$). We find very similar dependence relationships when the data are split into seasons. In the Supplementary Material, we show scatter plots of each

covariate against the response variable; these demonstrate a highly non-linear relationship for each explanatory variable with Y.

Next, we explore temporal relationships for the response variable Y. We first find temporal non-stationarity as the distribution of Y varies significantly with V_5 (season); see the Supplementary Material for more detail. The mean and range of Y is higher in season 1 than season 2, with greater extreme values observed in season 1. However, within each season, across months, there is little temporal variation in the distribution of Y. We also find that Y exhibits temporal independence at all lags, with auto-correlation function (acf) values close to zero; see the Supplementary Material.

As noted in Rohrbeck et al. (2023), 11.7% of the observations have at least one value missing completely at random (MCAR). A detailed breakdown of the pattern of missing predictor observations is provided in the Supplementary Material. Since we can assume the data are MCAR, ignoring the observations that have a missing predictor covariate will not bias our inference, however, a complete case analysis is undesirable due to the amount of data loss. To mitigate against this, we attempted to impute the observations where predictors are missing but ultimately could not find an imputation method that satisfactorily retained the dependence structure between the response and covariates, particularly in the tails of the distribution. Therefore, we use a case analysis approach, whereby an observation is only removed if a predictor covariate of interest is missing. This results in only 4% of observations being removed for our final model.

5.3.2 Methods

This section details our proposed method for modelling $Y \mid \boldsymbol{X}$. We start by outlining our model, which, due to the complex nature of the data, is a non-stationary GPD formulated with GAMs. While alternatives to GAMs, such as tensor models, could be appropriate for this type of data, these were not explored. To select a threshold for a non-stationary, covariate-dependent GPD model, we extend the method proposed by Murphy et al. (2024). We then outline our framework for inference and our model selection criteria based on k-fold cross-validation. Note that the same model formulation is used for both C1 and C2, with a small adjustment to the parameter estimation procedure for C2 to incorporate the loss function given in (5.3.2). Finally, we utilise equation (5.3.1) to obtain the marginal distribution of Y.

General model formulation

Let $\tilde{\boldsymbol{X}}_t$ denote the set of predictor covariates with $t \in \{1, \ldots, n\}$. Then y_t and $\tilde{\boldsymbol{x}}_t$ denote the observations of the response variable and predictive covariates, respectively. We consider models with the following form,

$$F_{Y_t|\tilde{\boldsymbol{X}}_t}(y_t|\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{x}}_t) = 1 - \zeta(\tilde{\boldsymbol{x}}_t)\left[1 + \xi(\tilde{\boldsymbol{x}}_t)\left(\frac{y_t - v(\tilde{\boldsymbol{x}}_t)}{\sigma(\tilde{\boldsymbol{x}}_t)}\right)\right]_+^{-1/\xi(\tilde{\boldsymbol{x}}_t)}, \tag{5.3.3}$$

where $v(\tilde{\boldsymbol{x}}_t)$ and $\zeta(\tilde{\boldsymbol{x}}_t)$ are a covariate-dependent threshold and rate parameter, respectively, such that the rate parameter corresponds to the probability of exceeding the threshold.

Our analysis in Section 5.3.1 indicates that V_3, V_5 (season), and V_6 (wind speed) exhibit non-trivial dependence relationships with the response variable. Therefore we assume these variables can be used as predictor variables for modelling Y, and set $\tilde{\boldsymbol{x}} := (\boldsymbol{V}_j)_{j\in\{3,5,6\}}$. Although V_7 (wind direction) also exhibits strong dependence with Y, we do not consider it here since it is highly correlated with wind speed so would involve adding complex interaction terms to the model formulation, and V_6 has a stronger relationship with Y compared to V_7 (see Figure 5.1).

Owing to the complex covariate structure observed in the data, as described in Section 5.3.1, we employ the flexible EVGAM framework proposed in Youngman (2019) for modelling tail behaviour. Under this framework, GAM formulations are used to capture non-stationarity in the threshold, scale and shape functions given in equation (5.3.3). Without loss of generality, consider the scale function $\sigma(\cdot)$. We assume that

$$h(\sigma(\tilde{\boldsymbol{x}})) = \psi_\sigma(\tilde{\boldsymbol{x}}), \quad \text{with} \quad \psi_\sigma(\tilde{\boldsymbol{x}}) = \beta_0 + \sum_{\kappa=1}^{K}\sum_{p=1}^{P_\kappa}\beta_{\kappa p}b_{\kappa p}(\tilde{\boldsymbol{x}}), \tag{5.3.4}$$

where $h(x) := \log(x)$ denotes the link function which ensures the correct support, with coefficients $\beta_0, \beta_{\kappa p} \in \mathbb{R}$ and basis functions $b_{\kappa p}$ for $p \in \{1, \ldots, P_\kappa\}, \kappa \in \{1, \ldots, K\}$, where K is the number of splines in the GAM formulation and P_κ is the basis dimension relating to spline κ. The basis functions can be in terms of individual covariates, i.e., $b_{\kappa p} : \mathbb{R} \mapsto \mathbb{R}$, or multiple covariates, i.e., $b_{\kappa p} : \mathbb{R}^m \mapsto \mathbb{R}$, $1 < m \leq 8$. Analogous forms can be taken for $v(\cdot)$ and $\xi(\cdot)$, adjusting the link function $h(\cdot)$ as appropriate, although these are not considered here for reasons detailed below. For clarity, the multiplication of the basis coefficients and functions is applied elementwise. If interactions between explanatory variables are considered, although that will not be the case here, then the Kronecker product would need to be considered.

To select an appropriate threshold, we employ the threshold selection method of Murphy et al. (2024) and extend this approach to select a threshold for non-stationary, covariate-dependent GPD models. The method selects a threshold based on minimising the expected

quantile discrepancy (EQD) between the sample quantiles and fitted GPD model quantiles. When fitting a non-stationary model, the excesses will not be identically distributed across covariates. Thus, to utilise the EQD method in this case, we use the fitted non-stationary GPD parameter estimates to transform the excesses to common standard exponential margins and compare sample quantiles against theoretical quantiles from the standard exponential distribution. This transformation is a common approach for checking the model fit of a non-stationary GPD (Coles, 2001).

We use a stepped-threshold according to season as there is clear variation in the distribution, and thereby the extremes, of Y between seasons; see the Supplementary Material for more details. Specifically, we set $v(\tilde{\boldsymbol{x}}_t) := \mathbb{1}(\tilde{x}_{2,t} = 1)v_1 + \mathbb{1}(\tilde{x}_{2,t} = 2)v_2$, $v_1, v_2 \in \mathbb{R}$, with corresponding rate parameter $\zeta(\tilde{\boldsymbol{x}}_t) := \mathbb{1}(\tilde{x}_{2,t} = 1)\zeta_1 + \mathbb{1}(\tilde{x}_{2,t} = 2)\zeta_2$, where $\zeta_1, \zeta_2 \in [0, 1]$ denote the probabilities of exceeding the threshold for seasons 1 and 2, respectively, and $\tilde{x}_{r,t}$ are realisations of the r^{th} component of $\tilde{\boldsymbol{x}}$ for $r \in \{1, 2, 3\}$. This seasonal threshold significantly improves model fits; see the Supplementary Material for further details. GAM forms for the threshold were also explored, but did not offer significant improvement. Furthermore, the smooth GAM formulation of the GPD scale parameter adequately captures any residual variation in the response arising due to covariate dependence.

Inference

For all GAM formulations, we only consider basis functions of singular explanatory variables, since specifying basis functions of multiple variables requires a detailed understanding of covariate interactions and can significantly increase the computational complexity of the modelling procedure (Wood, 2017). We keep the shape function $\xi(\boldsymbol{x}) := \xi \in \mathbb{R}$ constant across covariates; this is common in non-stationary analyses, since this parameter is difficult to estimate (Chavez-Demoulin and Davison, 2005). Within the GAM formulation, we consider several parametric forms to account for the predictive covariates in the scale parameter using linear models, indicator functions and splines.

When using splines, we are required to select a basis dimension $P_\kappa \in \mathbb{N}$; this determines the number of coefficients to be estimated. While ensuring the correct explanatory variables are included in the model is paramount, choosing the basis dimension is also a critical choice within spline modelling procedures (Wood, 2017) as it directly corresponds with the flexibility of the framework. If the basis dimension is chosen too high, then the splines will overfit to the data. Conversely, if the basis dimension is chosen too low, then the splines will not be flexible enough to capture the non-linear relationship with the response. We only consider splines for V_3 and V_6. For each $\tilde{X}_r$, $r \in \{1, 3\}$, we determine the basis dimension P_1 and P_2,

respectively, by first building a model for $Y_t \mid \tilde{X}_{r,t}$, to allow us to consider the effect of this predictor on the response directly. We vary the basis dimension and compare the resulting models using cross validation (CV), detailed in the following section. We set $P_1 = 4$ and $P_2 = 3$ for V_3 and V_6, respectively.

For C2, we incorporate the loss function of equation (5.3.2) into the estimation procedure. Let $\mathcal{I}_v := \{t \in \{1, \ldots, n\} \mid y_t > v(\tilde{\boldsymbol{x}}_t)\}$ denote the set of temporal indices corresponding to threshold exceedances and $n_v := |\mathcal{I}_v|$. We consider the objective function

$$S(\boldsymbol{\theta}) := -l_R(\boldsymbol{\theta}) + \sum_{i \in \mathcal{I}_v} \mathcal{L}(q_i, q_i^*; \boldsymbol{\theta})/n_v, \tag{5.3.5}$$

where $l_R(\boldsymbol{\theta})$ denotes the penalised log-likelihood function of the restricted maximum likelihood estimation (REML) approach (Wood, 2017), $\boldsymbol{\theta}$ denotes the parameter vector associated with the GPD formulation of equation (5.3.4), and $\sum_{i \in \mathcal{I}_v} \mathcal{L}(q_i, q_i^*; \boldsymbol{\theta})/n_v$ denotes the average loss between the sample quantiles of the transformed excesses and the theoretical standard exponential quantiles. Specifically, we transform the excesses, $(y_t - v(\tilde{\boldsymbol{x}}_t))_{t \in \mathcal{I}_v}$, to standard exponential margins using the non-stationary GPD parameter estimates $\boldsymbol{\theta}$ and compare the ordered excesses, $\boldsymbol{q}^*$, to the theoretical quantiles, $\boldsymbol{q}$, from a standard exponential distribution evaluated at probabilities $\{p_i = i/(n_v + 1), i = 1, \ldots, n_v\}$. Minimising the objective function $S(\boldsymbol{\theta})$ ensures that the parameter estimates also account for the loss function, $\mathcal{L}$. We use this formulation to adjust the GPD parameters for challenge C2 once a threshold is selected.

Model selection

To determine the best-fitting model, we use a forward selection process and aim to minimise the model's CV score. For each model, we apply k-fold CV (Hastie et al., 2001, Ch 7.) utilising the continuous ranked probability score (CRPS, Gneiting and Katzfuss, 2014) as our goodness-of-fit metric. CRPS describes the discrepancy between the predicted distribution function and observed values without the specification of empirical quantiles. We explore model ranking by taking both $k = 10$ and 50, and find that both give an equivalent ranking; we present results for the latter. We also provide the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values to aid in model selection. A subset of models used in the forward selection process are detailed in Table 5.1 where, for each model, we provide the change in the CRPS, AIC and BIC relative to model 1. The parameterisation of model 7 achieves the largest reduction for all three metrics relative to the baseline model.

Table 5.1: Table of selected models considered for challenge C1. $\mathbb{1}(\cdot)$ denotes an indicator function, $s_i(\cdot)$ for $i \in \{1, 2\}$ denote thin-plate regression splines, β_0, β_1 are coefficients to be estimated, and $\tilde{x}_{r,t}$ is defined as in the text. All values have been given to one decimal place.

Model	$\sigma(\tilde{\boldsymbol{x}}_t)$	ΔCRPS	ΔAIC	ΔBIC
1	β_0	0	0	0
2	$\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1)$	-0.5	-33.4	-26.1
3	$\beta_0 + s_1(\tilde{x}_{1,t})$	-0.9	-408.5	-379.2
4	$\beta_0 + s_2(\tilde{x}_{3,t})$	-0.5	-284.3	-276.8
5	$\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1) + s_1(\tilde{x}_{1,t})$	-0.9	-425.8	-388.1
6	$\beta_0 + s_1(\tilde{x}_{1,t}) + s_2(\tilde{x}_{3,t})$	-1.0	-752.7	-717.2
7	$\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1) + s_1(\tilde{x}_{1,t}) + s_2(\tilde{x}_{3,t})$	**-1.1**	**-780.0**	**-735.3**

5.3.3 Uncertainty

For each of the 100 different covariate combinations, $\tilde{\boldsymbol{x}}_i$ for $i \in \{1, \ldots, 100\}$, we need to construct central 50% confidence intervals. We use a bootstrapping procedure to avoid making potentially inaccurate assumptions, such as the asymptotic normality approximation of the penalised maximum likelihood estimates, for example. Traditional bootstrap approaches are non-parametric and randomly resample the data with replacement. However, in Section 5.3.1 we find that the response variable is dependent on covariates, and these covariates exhibit temporal dependence. A standard bootstrap procedure would therefore not retain this dependence. Instead, we preserve the temporal dependence structure of covariates and their relationship with the response variable by approximating our confidence intervals using the stationary, semi-parametric bootstrapping procedure adopted by D'Arcy et al. (2023).

First, the response variable Y_t is transformed to Uniform(0,1) margins to preserve its non-stationary behaviour; denote this sequence $U_t^Y = F_{Y_t|\tilde{\boldsymbol{X}}_t}(Y_t|\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{x}}_t)$ where $F_{Y_t|\tilde{\boldsymbol{X}}_t}$ is the estimated model given in equation (5.3.3). We then adopt the stationary bootstrap procedure of Politis and Romano (1994) to retain the temporal dependence in the response and explanatory variables by sampling blocks of consecutive observations. The block length L is random and simulated from a Geometric$(1/l)$ distribution, where the mean block length $l \in \mathbb{N}$ is carefully selected based on the autocorrelation function. This was selected at 50 days, the maximum lag for which the autocorrelation was significant across all variables; see the Supplementary Material. Denote this bootstrapped sequence on Uniform margins by U_t^B. We transform U_t^B back to the original scale using our fitted model, preserving the original structure of Y_t; we denote this series Y_t^B. Then we fit our model to Y_t^B to re-estimate all of the parameters and thus the quantile of interest. We repeat this procedure to obtain 200 bootstrap samples.

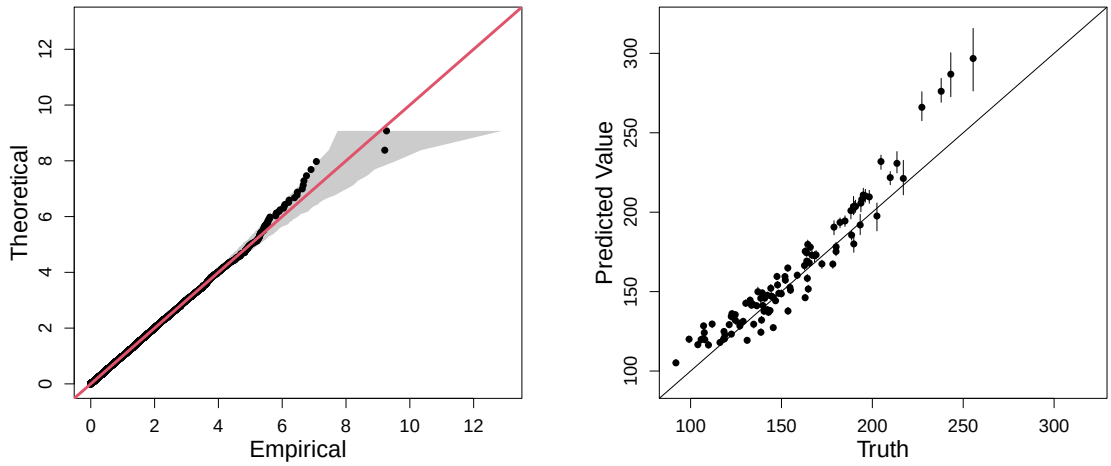

Figure 5.2: QQ plot for our final model (model 7 in Table 5.1) on standard exponential margins. The $y = x$ line is given in red, and the grey region represents the 95% tolerance bounds (left). Predicted 0.9999−quantiles against true quantiles for the 100 covariate combinations. The points are the median predicted quantile over 200 bootstrapped samples, and the vertical error bars are the corresponding 50% confidence intervals. The $y = x$ line is also shown (right).

5.3.4 Results

For C1, we use our final model of Section 5.3.2 to estimate the 0.9999-quantile of $Y \mid \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_i$, $i \in \{1, \ldots, 100\}$, for the set of 100 covariate combinations. The left panel of Figure 5.2 shows the quantile-quantile (QQ) plot for our model. There is general alignment between the model and empirical quantiles; however, there is some over-estimation in the upper tail, and our 95% tolerance bounds do not contain some of the most extreme response values. The right panel of Figure 5.2 shows our predicted quantiles, and their associated confidence intervals, compared to their true quantiles. As expected, our predictions tend to over-estimate the true quantiles. We note this figure is different from the one presented by Rohrbeck et al. (2023) due to an error in our code being fixed after submission. In this scenario, our estimated confidence intervals lead to a 14% coverage of the true quantiles, which does not alter our ranking for this challenge. Our performance and model improvements are discussed in Section 5.6.

For challenge C2, we estimate the quantile of interest as $\hat{q} = 213.1$ $(209.3, 242.1)$. A 95% confidence interval for the estimate is given in parentheses based on the bootstrapping procedure outlined in Section 5.3.2. Due to a coding error, this value differs from the original estimate submitted for the EVA (2023) Conference Data Challenge. The updated value over-estimates

compared to the truth ($q = 196.6$).

5.4 Challenge C3

5.4.1 Exploratory data analysis

For challenge C3, we are provided with 70 years of daily data of an environmental variable for three towns on the island of Coputopia. These data are denoted by $Y_{i,t}$, $i \in \{1, 2, 3\}$, $t \in \{1, \ldots, n\}$, where i is the index of each town and t is the point in time. Each year consists of 12 months, each lasting 25 days, resulting in $n = 21,000$ observations for each location.

We are also provided with daily covariate observations $\boldsymbol{X}_t = (S_t, A_t)$, where S_t and A_t denote seasonal and atmospheric conditions, respectively. Season is a binary variable, taking values in the set $\{1, 2\}$, with each year of observations exhibiting both seasons for exactly 150 consecutive days. Atmospheric conditions are piecewise constant over months, with large variation in the observed values between months. A descriptive figure of both covariates is given in the Supplementary Material.

In Rohrbeck et al. (2023), we are informed that $Y_{i,t}$ are distributed identically across all sites and over time, with standard Gumbel margins. However, it is not known whether the covariates $\boldsymbol{X}_t$ influence the dependence structure of $\boldsymbol{Y}_t := (Y_{1,t}, Y_{2,t}, Y_{3,t})$. We are also informed that, conditioned on covariates, the process is independent over time, i.e., $(\boldsymbol{Y}_t \mid \boldsymbol{X}_t) \perp\!\!\!\perp (\boldsymbol{Y}_{t'} \mid \boldsymbol{X}_{t'})$ for any $t \neq t'$. In this section, we examine what influence, if any, the covariate process $\boldsymbol{X}_t$ may have on the dependence structure of $\boldsymbol{Y}_t$.

We begin by transforming the time series $Y_{i,t}$ to standard exponential margins, denoted by $\boldsymbol{Z}_{i,t}$, via the probability integral transform. This transformation is common in the study of multivariate extremes and can simplify the description of extremal dependence (Keef et al., 2013a). To explore the extremal dependence in the Coputopia time series, we consider all 2- and 3-dimensional subvectors of the process, i.e., $\{Z_{i,t}, i \in I, t \in \{1, \ldots, n\}\}$, $I \in \mathcal{I} := \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. This separation is important to ensure the overall dependence structure is fully understood, since intermediate scenarios can exist where a random vector exhibits $\chi = 0$, but $\chi > 0$ for some 2-dimensional subvector(s) (Simpson et al., 2020).

Furthermore, to explore the impact of covariates on the dependence structure, we partition the time series into subsets using the covariates. For the seasonal covariate, let $G^S_{I,j} := \{Z_{i,t}, i \in I, S_t = j\}$ for $j = 1, 2$, and for the atmospheric covariate, let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ denote the permutation associated with the order statistics of A_t, defined so that

ties in the data are accounted for. We then split the data into 10 equally sized subsets corresponding to the atmospheric order statistics, i.e., $G^A_{I,k} := \left\{Z_{i,t}, i \in I, t \in \Sigma^k\right\}$ for $k = 1, 2, \ldots, 10$, where $\Sigma^k := \{t \mid (k-1)n/10+1 \leq \pi(t) \leq kn/10\}$. Thus, the atmospheric values associated with each subset $G^A_{I,k}$ will increase over k.

The idea behind these subsets is to examine whether altering the values of either covariate impacts the extremal dependence structure. Consequently, we set $u = 0.9$ and estimate $\chi(u)$ using the techniques outlined in Section 5.2, with uncertainty quantified through bootstrapping with 200 samples. The bootstrapped χ estimates for $G^A_{I,k}$ with $I = \{1, 2, 3\}$ are given in Figure 5.3. The plots for the remaining index sets in $\mathcal{I}$, along with the subsets associated with the seasonal covariate, are given in the Supplementary Material. The estimates of χ appear to vary, in the majority of cases, across both subset types (seasonal and atmospheric), suggesting both covariates have an impact on the dependence structure. For the atmospheric process in particular, the values of χ tend to decrease for higher atmospheric values, suggesting a negative association between the strength of positive extremal dependence and the atmospheric covariate. We also observe that across all subsets, χ appears consistently low in magnitude, suggesting the extremes of some, if not all, of the sub-vectors are unlikely to occur simultaneously. As such, for modelling the Coputopia time series, we require a framework that can capture such forms of dependence. We also consider pointwise estimates of the function $\lambda(\cdot)$, as defined later in equation (5.4.1), over $G^S_{I,j}$ and $G^A_{I,k}$ for fixed simplex points; these results are given in the Supplementary Material. Similar to χ, estimates of $\lambda(\cdot)$ vary significantly across subsets, providing additional evidence of non-stationarity within extremal dependence structure.

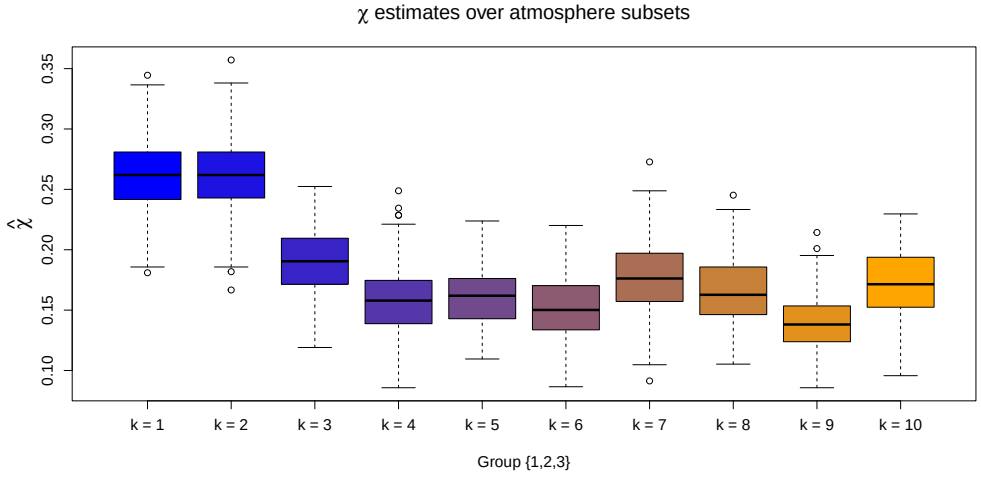

Figure 5.3: Boxplots of empirical χ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

5.4.2 Modelling of joint tail probabilities under asymptotic independence

For challenge C3, we are required to estimate probabilities $p_1 := \Pr(Y_1 > y, Y_2 > y, Y_3 > y)$ and $p_2 := \Pr(Y_1 > v, Y_2 > v, Y_3 < m)$, with $y = 6$, $v = 7$ and $m = -\log(\log(2))$. Note that p_1 and p_2 are independent of the covariate process and correspond to different extremal regions in $\mathbb{R}^3$; we refer to p_1 and p_2 as parts 1 and 2 of the challenge, respectively. For the remainder of this section we will consider the transformed exponential variables (Z_1, Z_2, Z_3), omitting the subscript t for ease of notation. Observe that $F_{(-Z_3)}(z) = e^z$, for $z < 0$; setting $\tilde{Z}_3 := -\log\big(1 - \exp(-Z_3)\big)$, we have

$$p_2 = \Pr(Z_1 > \tilde{v}, Z_2 > \tilde{v}, Z_3 < \tilde{m}) = \Pr\Big(Z_1 > \tilde{v}, Z_2 > \tilde{v}, \tilde{Z}_3 > \tilde{m}\Big),$$

where $\tilde{v}$ and $\tilde{m}$ denote the values v and m transformed to the standard exponential scale, e.g., $\tilde{v} := -\log\big(1 - \exp(-\exp(-v))\big)$. Similarly, we have $p_1 = \Pr(Z_1 > \tilde{y}, Z_2 > \tilde{y}, Z_3 > \tilde{y})$. Consequently, both p_1 and p_2 can be considered as joint survivor probabilities.

Since not all extremes of Z_1, Z_2 and Z_3 are observed simultaneously, we employ the framework by Wadsworth and Tawn (2013), which is a generalisation of the approach proposed in Ledford and Tawn (1996). The model of Wadsworth and Tawn (2013) assumes that for any ray $\boldsymbol{\omega} \in \boldsymbol{S}^2 := \big\{(\omega_1, \omega_2, \omega_3) \in [0,1]^3 : \omega_1 + \omega_2 + \omega_3 = 1\big\}$, where $\boldsymbol{S}^2$ denotes the standard 2-dimensional simplex,

$$\begin{aligned} \Pr(Z_1 > \omega_1 r,\, Z_2 > \omega_2 r,\, Z_3 > \omega_3 r) &= \Pr\big(\min\{Z_1/\omega_1,\, Z_2/\omega_2,\, Z_3/\omega_3\} > r\big) \\ &= \mathcal{L}(e^r; \boldsymbol{\omega}) e^{-r\lambda(\boldsymbol{\omega})}, \end{aligned} \tag{5.4.1}$$

as $r \to \infty$, where $\lambda(\boldsymbol{\omega}) \geq \max(\boldsymbol{\omega})$ is known as the angular dependence function (ADF). Asymptotic dependence occurs at the lower bound, i.e., $\lambda(\boldsymbol{\omega}) = \max(\boldsymbol{\omega})$ for all $\boldsymbol{\omega} \in \boldsymbol{S}^2$, and the coefficient of tail dependence is related to the ADF via $\eta = 1/\{3\lambda(1/3, 1/3, 1/3)\}$. In practice, equation (5.4.1) can be used to evaluate extreme joint survivor probabilities; in particular, probabilities p_1 and p_2 can be identified with the rays $\boldsymbol{\omega}^{(1)} := (\tilde{y}, \tilde{y}, \tilde{y})/r^{(1)}$ and $\boldsymbol{\omega}^{(2)} := (\tilde{v}, \tilde{v}, \tilde{m})/r^{(2)}$ in $\boldsymbol{S}^2$, respectively, where $r^{(1)} := \tilde{y} + \tilde{y} + \tilde{y}$ and $r^{(2)} := \tilde{v} + \tilde{v} + \tilde{m}$. See Section 5.4.4 for further details.

5.4.3 Accounting for non-stationary dependence

In the stationary setting, pointwise estimates of $\lambda(\cdot)$ can be obtained via the Hill estimator (Hill, 1975), from which tail probabilities can be approximated. However, alternative procedures are required for data exhibiting trends in dependence, such as the Coputopia

data set. Existing approaches for capturing non-stationary dependence structures are sparse in the extremes literature, and most approaches are limited to asymptotically dependent data structures. For the case when data are not asymptotically dependent, Mhalla et al. (2019) and Murphy-Barltrop and Wadsworth (2024) propose non-stationary extensions of the Wadsworth and Tawn (2013) framework, while Jonathan et al. (2014) and Guerrero et al. (2023) propose non-stationary extensions of the Heffernan and Tawn (2004) model (see Murphy-Barltrop and Wadsworth (2024) for a detailed review).

To account for non-stationary dependence in C3, we propose an extension of the Wadsworth and Tawn (2013) framework. With $\boldsymbol{Z}_t = (Z_{1,t}, Z_{2,t}, Z_{3,t})$ and $\boldsymbol{X}_t$, defined as in Section 5.4.1, we define the structure variable $T_{\boldsymbol{\omega},t} := \min\{Z_{1,t}/\omega_1, Z_{2,t}/\omega_2, Z_{3,t}/\omega_3\}$, for any $\boldsymbol{\omega} \in \boldsymbol{S}^2$; we refer to $T_{\boldsymbol{\omega},t}$ as the min-projection variable at time t. From Section 5.4.1, we know that the joint distribution of $\boldsymbol{Z}_t$ is not identically distributed over t; which implies non-stationarity in the distribution of $T_{\boldsymbol{\omega},t}$. To account for this, Mhalla et al. (2019) and Murphy-Barltrop and Wadsworth (2024) assume the following model given the vector of covariates $\boldsymbol{x}_t$:

$$\Pr\left(T_{\boldsymbol{\omega},t} > u \mid \boldsymbol{X}_t = \boldsymbol{x}_t\right) = \mathcal{L}\left(e^u \mid \boldsymbol{\omega}, \boldsymbol{x}_t\right) e^{-\lambda(\boldsymbol{\omega};\boldsymbol{x}_t)u} \text{ as } u \to \infty, \tag{5.4.2}$$

for all t, where $\lambda\left(\cdot; \boldsymbol{x}_t\right)$ denotes the non-stationary ADF. Note that this assumption is very similar in form to equation (5.4.1), with the primary difference being the function $\lambda(\cdot; \boldsymbol{x}_t)$ is non-stationary over t. From equation (5.4.2), we have

$$\Pr\left(T_{\boldsymbol{\omega},t} - u > z \mid T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\right) = e^{-\lambda(\boldsymbol{\omega};\boldsymbol{x}_t)z} \text{ as } u \to \infty, \tag{5.4.3}$$

for $z > 0$. Consequently, equation (5.4.2) is equivalent to assuming $(T_{\boldsymbol{\omega},t} - u) \mid \{T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{Exp}(\lambda\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right))$ as $u \to \infty$.

We found that equation (5.4.2) was not flexible enough to capture the tail of $T_{\boldsymbol{\omega},t}$ for the Coputopia data; see Section 5.4.3 for further discussion. Thus, we propose the following model: given any $z > 0$ and a fixed $\boldsymbol{\omega} \in \boldsymbol{S}^2$, we assume

$$\Pr\left(T_{\boldsymbol{\omega},t} - u > z \mid T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\right) = \left(1 + \frac{\xi\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right) z}{\sigma\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right)}\right)^{-1/\xi(\boldsymbol{\omega};\boldsymbol{x}_t)} \text{ as } u \to \infty, \tag{5.4.4}$$

where $\sigma(\cdot; \boldsymbol{x}_t), \xi(\cdot; \boldsymbol{x}_t)$ are non-stationary scale and shape parameter functions, respectively. This is equivalent to assuming $(T_{\boldsymbol{\omega},t} - u) \mid \{T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{GPD}(\sigma\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right), \xi\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right))$ as $u \to \infty$, and equation (5.4.3) is recovered by taking the limit as $\xi\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right) \to 0$ for all t.

Our proposed formulation in equation (5.4.4) allows for additional flexibility within the modelling framework by including a GPD shape parameter $\xi\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right)$, which quantifies the tail

behaviour of $T_{\boldsymbol{\omega},t}$. Given the wide range of distributions in the domain of attraction of a GPD (Pickands III, 1975), it is reasonable to assume that the tail of $T_{\boldsymbol{\omega},t}$ can be approximated by equation (5.4.4). For the Coputopia time series, this assumption appears valid, as demonstrated by the diagnostics in Section 5.4.3.

Model fitting

To apply equation (5.4.4), we first fix $\boldsymbol{\omega} \in \boldsymbol{S}^2$ and assume that the formulation holds approximately for some sufficiently high threshold level from the distribution of $T_{\boldsymbol{\omega},t}$; we denote the corresponding quantile level by $\tau \in (0,1)$. For simplicity, the same quantile level is considered across all t. Further, let $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)$ denote the corresponding threshold function, i.e., $\Pr(T_{\boldsymbol{\omega},t} \leq v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t) \mid \boldsymbol{X}_t = \boldsymbol{x}_t) = \tau$ for all t. Under our assumption, we have $(T_{\boldsymbol{\omega},t} - v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) \mid \{T_{\boldsymbol{\omega},t} > v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t), \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{GPD}(\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t), \xi(\boldsymbol{\omega}; \boldsymbol{x}_t))$. We emphasise that $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)$ is not constant in t, and we would generally expect $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t) \neq v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_{t'})$ for $t \neq t'$.

As detailed in Section 5.4.2, both p_1 and p_2 can be associated with points on the simplex $\boldsymbol{S}^2$, denoted by $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$, respectively. Letting $\boldsymbol{\omega} \in \{\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}\}$, our estimation procedure consists of two stages: estimation of the threshold function $v_\tau(\boldsymbol{\omega}, \boldsymbol{z}_t)$ for a fixed $\tau \in (0,1)$, followed by estimation of GPD parameter functions $\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t), \xi(\boldsymbol{\omega}; \boldsymbol{x}_t)$. For both steps, we take a similar approach to Section 5.3.2 and use GAMs to capture these covariate relationships. To simplify our approach, we falsely assume that the atmospheric covariate A_t is continuous over t; this step allows us to utilise GAM formulations containing smooth basis functions. Given the significant variability in A_t between months, discrete formulations for this covariate would significantly increase the number of model parameters and result in higher variability.

Let $\log(v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) = \psi_v(\boldsymbol{x}_t)$, $\log(\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t)) = \psi_\sigma(\boldsymbol{x}_t)$ and $\xi(\boldsymbol{\omega}; \boldsymbol{x}_t) = \psi_\xi(\boldsymbol{x}_t)$ denote the GAM formulations of each function, where ψ_- denotes the basis representation of equation (5.3.4). Exact forms of basis functions are specified in Section 5.4.3. As in Section 5.3.2, model fitting is carried out using the `evgam` software package (Youngman, 2022). For the first stage, $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)$ is estimated by exploiting a link between the loss function typically used for quantile regression and the asymmetric Laplace distribution (Yu and Moyeed, 2001). The spline coefficients associated with ψ_σ and ψ_ξ are estimated subsequently using the obtained threshold exceedances.

Selection of GAM formulations and diagnostics

Prior to estimation of the threshold and parameter functions, we specify a quantile level τ and formulations for each of the GAMs. To begin, we fix $\tau = 0.9$ and consider a variety of formulations for each ψ_v, ψ_σ and ψ_ξ. By comparing metrics for model selection, namely AIC, BIC and CRPS, we found the following formulations to be sufficient

$$\psi_v(\boldsymbol{x}_t) = \beta_u + s_v(a_t) + \beta_s \mathbb{1}(s_t = 2), \quad \psi_\sigma(\boldsymbol{x}_t) = \beta_\sigma + s_\sigma(a_t) \quad \text{and} \quad \psi_\xi(\boldsymbol{x}_t) = \beta_\xi, \tag{5.4.5}$$

for parts 1 and 2, where $\beta_u, \beta_\sigma, \beta_\xi \in \mathbb{R}$ denote constant intercept terms, $\mathbb{1}$ denotes the indicator function with corresponding coefficient $\beta_s \in \mathbb{R}$, and s_u, s_σ denote cubic regression splines of dimension 10. The shape parameter is set to constant for the reasons outlined in Section 5.3.2. Cubic basis functions are used for ψ_v and ψ_σ since they have several desirable properties, including continuity and smoothness (Wood, 2017). A dimension of size 10 appears more than sufficient to capture the trends relating to the atmosphere variable. Alternative formulations were tested for both parts, but this made little difference to the resulting model fits.

We remark that the seasonal covariate is only present with the formulation for ψ_v. Once accounted for in the non-stationary threshold, the seasonal covariate appeared to have little influence on the fitted GPD parameters. More complex GAM formulations were tested involving interaction terms between the seasonal and atmospheric covariates, which showed little to no improvement in model fits. Thus, we prefer the simpler formulations on the basis of parsimony.

With GAM formulations selected, we now consider the quantile level $\tau \in (0, 1)$. To assess sensitivity in our formulation, we set $\mathrm{T} := \{0.8, 0.81, \ldots, 0.99\}$ and fit the GAMs outlined in equation (5.4.5) for each $\tau \in \mathrm{T}$. Letting $\delta_{\boldsymbol{\omega},t}$ and $\mathcal{T}_\tau := \{t \in \{1, \ldots, n\} \mid \delta_{\boldsymbol{\omega},t} > v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)\}$ denote the min-projection observations and indices of threshold-exceeding observations, respectively, we expect the set $\mathcal{E} := \{-\log\left\{1 - F_{GPD}(\delta_{\boldsymbol{\omega},t} - v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) \mid \sigma\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right), \xi\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right)\right\} \mid t \in \mathcal{T}_\tau\}$ to follow a standard exponential distribution.

With all exceedances transformed to a unified scale, we compare the empirical and model exponential quantiles using QQ plots, through which we assess the relative performance of each $\tau \in \mathrm{T}$. We selected τ values for which the empirical and theoretical quantiles appeared most similar in magnitude. From this analysis, we set $\tau = 0.83$ and $\tau = 0.85$ for parts 1 and 2, respectively. The corresponding QQ plots are given in Figure 5.4, where we observe reasonable agreement between the empirical and theoretical quantiles. However, whilst these values appeared optimal within T, we stress that adequate model fits were also obtained for other quantile levels, suggesting our modelling procedure is not particularly sensitive to the

exact choice of quantile. Furthermore, we also tested a range of quantile levels below the 0.8-level, but were unable to improve the quality of model fits.

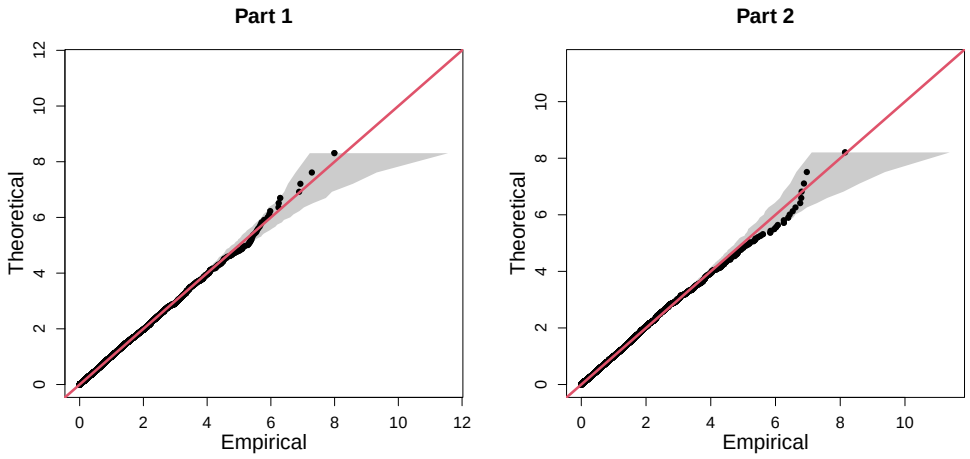

Figure 5.4: Final QQ plots for parts 1 (left) and 2 (right) of C3, with the $y = x$ line given in red. In both cases, the grey regions represent the 95% bootstrapped tolerance bounds.

Plots illustrating the estimated GPD scale parameter functions are given in the Supplementary Material, with the resulting dependence trends in agreement with the observed trends from Section 5.4.1. We also remark that the estimated GPD shape parameters obtained for parts 1 and 2 were $0.042\,(0.01, 0.075)$ and $0.094\,(0.059, 0.128)$, respectively, where the brackets denote 95% confidence intervals obtained using posterior sampling (Wood, 2017). These estimates, which indicate slightly heavy-tailed behaviour within the min-projection variable, provide insight into why the original exponential modelling framework is not appropriate for C3.

Overall, these results suggest different extremal dependence trends exist for the two simplex points $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$, illustrating the importance of the flexibility in our model. These findings are also in agreement with empirical trends observed in Section 5.4.1, suggesting our modelling framework is successfully capturing the underlying extremal dependence structures.

5.4.4 Results

Given estimates of threshold and parameter functions, probability estimates can be obtained via Monte Carlo techniques. Taking p_1, for instance, we have

$$\begin{aligned} p_1 &= \Pr(Z_1 > \tilde{y}, Z_2 > \tilde{y}, Z_3 > \tilde{y}) \\ &= \Pr\left(\min\left(Z_1/\omega_1^{(1)}, Z_2/\omega_2^{(1)}, Z_3/\omega_3^{(1)}\right) > r^{(1)}\right) \\ &= \int_{\boldsymbol{X}_t} \Pr\left(T_{\boldsymbol{\omega}^{(1)},\,t} > r^{(1)} \mid \boldsymbol{X}_t = \boldsymbol{x}_t\right) f_{\boldsymbol{X}_t}(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \\ &= (1-\tau)\int_{\boldsymbol{X}_t} \Pr(T_{\boldsymbol{\omega}^{(1)},\,t} > r^{(1)} \mid T_{\boldsymbol{\omega}^{(1)},\,t} > v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t), \boldsymbol{X}_t = \boldsymbol{x}_t) f_{\boldsymbol{X}_t}(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \\ &\approx \frac{1-\tau}{n}\sum_{t=1}^{n}\left(1 + \frac{\xi(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t)\left(r^{(1)} - v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t)\right)}{\sigma\left(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t\right)}\right)^{-1/\xi\left(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t\right)}, \end{aligned}$$

assuming $\{\boldsymbol{x}_t : t \in \{1, \ldots, n\}\}$ is a representative sample from $\boldsymbol{X}_t$. The procedure for p_2 is analogous. We note that this estimation procedure is only valid when $r^{(1)} > v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t)$, or $r^{(2)} > v_\tau(\boldsymbol{\omega}^{(2)}, \boldsymbol{x}_t)$, for all t: however, for each $\tau \in \mathrm{T}$, this inequality is always satisfied, owing to the very extreme nature of the probabilities in question. Through this approximation, we obtain $\hat{p}_1 = 1.480 \times 10^{-5}$ and $\hat{p}_2 = 2.461 \times 10^{-5}$.

5.5 Challenge C4

5.5.1 Exploratory data analysis

Challenge C4 entails estimating survival probabilities across 50 locations on the island of Utopula. As stated in Rohrbeck et al. (2023), the Utopula island is split in two administrative areas, for which the respective regional governments 1 and 2 have collected data concerning the variables $Y_{i,t}$, $i \in I = \{1, \ldots, 50\}$, $t \in \{1, \ldots, 10,000\}$. Index i denotes the i^{th} location, with locations $i \in \{1, \ldots, 25\}$ and $i \in \{26, \ldots, 50\}$ belonging to the administrative areas of governments 1 and 2, respectively. Index t denotes the time point in days; however, since $Y_{i,t}$ are IID for all i, we drop the subscript t for the remainder of this section.

Since many multivariate extreme value models are only applicable in low-to-moderate dimensions, we consider dimension reduction based on an exploration of the extremal dependence structure of the data. In particular, we analyse pairwise estimates of the extremal dependence coefficient $\chi(u)$, introduced in equation (5.2.2), for all possible pairwise combinations of sites; the resulting estimates, using $u = 0.95$, are presented in the heat map of Figure 5.5. Identification of any dependence clusters is achieved through visual investigation, which seems

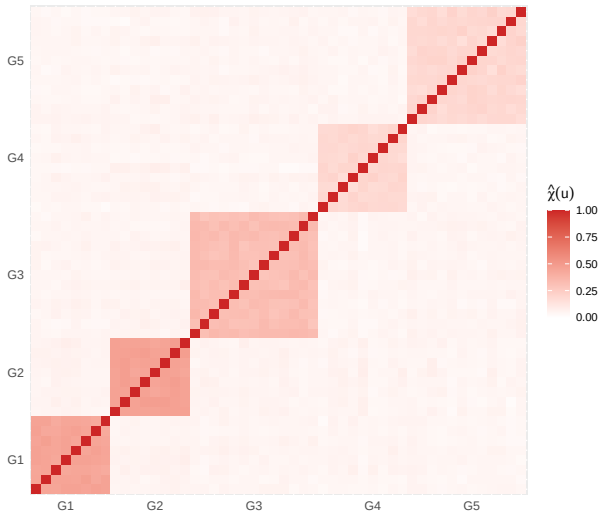

Figure 5.5: Heat map of estimated empirical pairwise $\chi(u)$ extremal dependence coefficients with $u = 0.95$.

appropriate for this data. We note, however, that should visual considerations not suffice, alternative more sophisticated clustering methods are available and can be applied; see for example Bernard et al. (2013).

Figure 5.5 suggests the existence of 5 distinct subgroups where all variables within each subgroup have similar extremal dependence characteristics, while variables in different subgroups appear to be approximately independent of each other in the extremes. It is worth mentioning that the same clusters are identified when we analyse pairwise estimates of the extremal dependence coefficient $\eta(u)$; the resulting estimates can be found in the Supplementary Material. Moreover, examining the magnitudes of $\chi(\cdot)$ and $\eta(\cdot)$ estimates, it does not appear reasonable to assume asymptotic dependence between variables in the same group. We therefore consider models that can be applied to data structures that do not take their extreme values simultaneously. The indices of the five aforementioned subgroups are $G_1 = \{4, 14, 19, 28, 30, 38, 43, 44\}$, $G_2 = \{3, 10, 15, 18, 22, 29, 45, 47\}$, $G_3 = \{8, 21, 25, 26, 32, 33, 34, 40, 41, 42, 48, 49, 50\}$, $G_4 = \{1, 2, 5, 7, 9, 17, 20, 31, 46\}$ and $G_5 = \{6, 11, 12, 13, 16, 23, 24, 27, 35, 36, 37, 39\}$. Groups G_1 and G_2 include the most strongly dependent variables (shown by the darkest color blocks in Figure 5.5), followed by group G_3, while groups G_4 and G_5 contain the most weakly dependent variables. We henceforth assume independence between these groups of variables, i.e., $\Pr((Y_i)_{i\in G_k} \in A_k, (Y_i)_{i\in G_{k'}} \in A_{k'}) = \Pr((Y_i)_{i\in G_k} \in A_k) \Pr((Y_i)_{i\in G_{k'}} \in A_{k'})$, $A_k \subset \mathbb{R}^{|G_k|}, A_{k'} \subset \mathbb{R}^{|G_{k'}|}$, for any $k \neq k' \in \{1, \ldots, 5\}$.

Challenge C4 requires us to estimate the probabilities $p_1 = \Pr(Y_i > s_i;\, i \in I)$ and $p_2 =$

$\Pr(Y_i > s_1; i \in I)$, where $s_i := \mathbb{1}(i \in \{1, 2, \ldots, 25\})s_1 + \mathbb{1}(i \in \{26, 27, \ldots, 50\})s_2$ and s_1 (s_2) denotes the marginal level exceeded once every year (month) on average. Under the assumption of independence between groups, the challenge can be broken down to 5 lower-dimensional challenges involving the estimation of joint tail probabilities for each G_k, $k \in \{1, \ldots, 5\}$. These can then be multiplied together to obtain the required overall probabilities due to (assumed) between-group independence; specifically, we have $p_1 = \prod_{k=1}^{5} \Pr(Y_i > s_i;\ i \in G_k)$ and $p_2 = \prod_{k=1}^{5} \Pr(Y_i > s_1;\ i \in G_k)$.

5.5.2 Conditional extremes

The conditional multivariate extreme value model (CMEVM) of Heffernan and Tawn (2004) provides a flexible multivariate extreme value framework capable of capturing a range of extremal dependence forms without making assumptions about the specific form of joint dependence structure. Consider a d-dimensional random variable $\boldsymbol{W} = (W_1, \ldots, W_d)$ on standard Laplace margins. For $i \in \{1, \ldots, d\}$, the CMEVM approach assumes the existence of parameter vectors $\boldsymbol{\alpha}_{-|i} \in [-1, 1]^{d-1}$ and $\boldsymbol{\beta}_{-|i} \in (-\infty, 1]^{d-1}$ such that

$$\lim_{u_i \to \infty} \Pr\left\{ \boldsymbol{W}_{-i} \leq \boldsymbol{\alpha}_{-|i} W_i + W_i^{\boldsymbol{\beta}_{-|i}} \boldsymbol{z}_{|i}, W_i - u_i > w \mid W_i > u_i \right\} = e^{-w} H_{|i}\left(\boldsymbol{z}_{|i}\right), \quad w > 0,$$

with non-degenerate distribution function $H_{|i}(\cdot)$, vector operations being applied componentwise, and conditional threshold u_i. The vector $\boldsymbol{W}_{-i}$ denotes $\boldsymbol{W}$ excluding its i^{th} component and $\boldsymbol{z}_{|i}$ is within the support of the residual random vector $\boldsymbol{Z}_{|i} = (\boldsymbol{W}_{-i} - \boldsymbol{\alpha}_{-|i} w_i)/w_i^{\boldsymbol{\beta}_{-|i}} \sim H_{|i}(\cdot)$. We apply this model to data where $W_i > u_i$, for some finite conditioning threshold u_i, to estimate the probabilities p_1 and p_2 defined in Section 5.5.1, using the inference procedure of Keef et al. (2013a).

5.5.3 Results

Let $\boldsymbol{W} := (W_1, \ldots, W_{50})$ denote the random vector after transformation to standard Laplace margins. This vector is divided into the five subgroups identified in Section 5.5.1, and the subgroup probabilities are estimated using predictions obtained from the sampling method of Heffernan and Tawn (2004). We condition on the first variable of each subgroup being extreme, and simulate 10^8 predictions from each of the resulting fitted conditional extremes models. To account for uncertainty in the estimates, we perform a parametric bootstrapping procedure with 100 samples.

Sensitivity analyses of the estimated probabilities to the choice of conditioning variable suggest no significant effect. Furthermore, we consider a range of conditioning thresholds;

the corresponding estimates of subgroup probabilities defined in Section 5.5.1 appear relatively stable with respect to the conditioning threshold quantile. We ultimately select 0.85-quantiles for the conditioning thresholds of our final probability estimates. These are given by $\hat{p}_1 = 1.094 \times 10^{-26}$ $(2.150 \times 10^{-36}, 1.359 \times 10^{-24})$ and $\hat{p}_2 = 1.076 \times 10^{-31}$ $(1.596 \times 10^{-46}, 1.850 \times 10^{-29})$, with 95% confidence intervals obtained from parametric bootstrapping given in parentheses.

5.6 Discussion

In this paper, we have proposed a range of statistical methods for estimating extreme quantities for challenges C1-C4. For the univariate challenge C1, we estimated the 0.9999-quantile, and the associated 50% confidence intervals, of $Y \mid \boldsymbol{X} = \boldsymbol{x_i}$, $i \in \{1, \ldots, n\}$. For challenge C2, we estimated a quantile, corresponding to a once in 200 year level, of the marginal distribution Y whilst incorporating the loss function in equation (5.3.2). Overall we ranked 6^{th} and and 4^{th} for challenges C1 and C2, respectively.

For challenge C1, our final model (model 7 in Table 5.1) was chosen to minimise the model selection criteria; however, QQ plots showed over-estimation of the most extreme values of the response (see Figure 5.2). As a result, the conditional quantiles calculated for C1 are generally over-estimated when compared with the true quantiles. If we ignored the model selection criteria and chose the model based on a visual assessment of QQ plots, we would have chosen model 5 in Table 5.1 and this would have covered the true quantile on fewer occasions than our chosen model. Therefore, the main issue with our results concerns the width of the confidence intervals.

Narrow confidence intervals are an indication of over-fitting and this could have arisen in several places. For instance, Rohrbeck et al. (2023) suggested all the seasonality is captured in the threshold, while our model includes a seasonal threshold and a covariate for seasonality in the scale parameter of the GPD model. As well as over-fitting, the model may not have been flexible enough; this could be, in part, due to our model missing covariates. For instance, the true model contained V_2 as a covariate (Rohrbeck et al., 2023) whilst our model did not. In addition, the basis dimensions for our splines are low. In practice, a higher dimension than we would expect should be considered and, although we chose the dimension using a model-based approach, it may have resulted in the splines not being flexible enough to capture all of the trends in the data.

Narrow confidence intervals may have also resulted from the choice of uncertainty quantification procedure. Changing the average block length l in our stationary bootstrap procedure

would alter the confidence interval widths, although this was carefully chosen to reflect the temporal dependence in the data. Alternative methods, such as the standard bootstrap procedure or the delta method, could be implemented to investigate how this affects the confidence interval widths. We expect that such confidence intervals will be wider than those presented here since the dependence in the data is not accounted for, but assuming temporal independence would be inaccurate. Therefore, whilst adopting an alternative procedure may widen confidence intervals, thus improving our performance, such intervals may not be well calibrated for this data set.

The over-fitting and over-estimation issues encountered in C1 are carried through to C2 since the same model is used for both challenges. However, one aspect specific to C2 is the choice of quantile evaluation within the loss function. Many methods exist for evaluating the non-stationary quantiles which feed into the loss function term of the objective function $S(\boldsymbol{\theta})$ in equation (5.3.5). As the loss function will be dominated by the log-likelihood in $S(\boldsymbol{\theta})$, we choose to transform to standard exponential margins when evaluating the quantiles in order to give more importance to the loss function. Since the data is light tailed ($\xi < 0$) this transformation elongates the tail and therefore inflates any deviations between the model and theoretical quantiles which in turn, inflates the contribution of the average loss function to $S(\boldsymbol{\theta})$. However, this approach means that the objective function will have a preference to minimise the deviations in the upper-tail of the distribution, leading to potential over-fitting to the upper-tail and possibly, a poor fit in the rest of the tail. This may not necessarily be undesirable since the loss function penalises under-estimation more than over-estimation, however, since the model in C1 already over-fits, this method may only exacerbate the problem for C2.

For the first multivariate challenge C3, we employed an extension of the method proposed by Wadsworth and Tawn (2013) to estimate probabilities of three variables lying in extremal sets. Our extension accounts for non-stationarity in the extremal dependence structure, with GAMs used to represent covariate relationships. The QQ plots for the resulting model suggested reasonable fits. For this challenge, we ranked 5^{th} and our estimates are on the same order of magnitude as the truth (Rohrbeck et al., 2023).

We note similarities in the methodologies presented for the challenges C1, C2, and C3. Specifically, each of the proposed methods used the EVGAM framework for capturing non-stationary tail behaviour via a generalised Pareto distribution. We acknowledge that the model selection tool proposed for C1 and C2 could also be applied for C3. However, we opted not to use this tool for several reasons. Firstly, unlike the univariate setting, there is no guarantee of convergence to a GPD in the limit, and the GPD tail assumption thereby

needs to be tested. Moreover, in exploratory analysis, we tested the model selection tool for C3 but found the selected models and quantiles to not be satisfactory, particularly in the upper tail of the min-projection variable. We therefore selected a model manually, using QQ plots to evaluate performance. Exploring threshold and model selection techniques for multivariate extremes represents an important area of research.

In the final multivariate challenge C4, we estimated very high-dimensional joint survival probabilities. To do so, we split the probability into 5 lower-dimensional components which are assumed independent of each other, then estimated each using the CMEVM of Heffernan and Tawn (2004). In the final rankings of Rohrbeck et al. (2023), we ranked 3$^{\text{rd}}$ for this challenge. A more prudent method could have been implemented, as groups of variables were never truly independent. Alternatively, although we achieve relatively stable probability estimates with respect to threshold in Section 5.2 (see Supplementary Material for details), our approach could potentially have been improved by estimating individual group probabilities across varying thresholds and taking an average value as our final result. We also do not report the effect of the choice of the conditioning variable on our estimates. Preliminary analysis suggested this to be negligible. However, conditioning on each site in a given subgroup and then taking a weighted sum of the resulting probabilities (e.g., Keef et al., 2013b) may have resulted in more robust estimates.

Declarations

Ethical Approval and Consent to Participate

Not Applicable

Consent for Publication

Not Applicable

Materials Availability

Materials that support the findings of this study are available from the corresponding author upon reasonable request.

Code Availability

Code supporting the findings of this study is available from the corresponding author upon reasonable request.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

This work was supported by EPSRC grant numbers EP/L015692/1, EP/S022252/1, EP/W523811/1 and EP/W524438/1, and SFI grant number 18/CRT/6049.

Authors' contributions

All authors contributed equally to this work.

Acknowledgments

This paper is based on work completed while Lídia André, Eleanor D'Arcy, Conor Murphy, Callum Murphy-Barltrop and Matthew Speers were part of the EPSRC funded STOR-i Centre for Doctoral Training (EP/L015692/1, EP/S022252/1), Ryan Campbell, Aiden Farrell and Lydia Kakampakou were part of EPSRC funded projects (EP/W523811/1, EP/W524438/1), and Dáire Healy was part of the Science Foundation Ireland funded project (18/CRT/6049). We are grateful to the two referees and editors for constructive comments and suggestions that have improved this article. We would also like to thank Ben Youngman for his assistance with the `evgam` package in the `R` computing language, as well as Christian Rohrbeck, Emma Simpson and Jonathan Tawn for their hard work in organising the data challenge.

Supplementary Information

Supplementary Material for "Extreme value methods for estimating rare events in Utopia": File containing additional figures. (supplementary.pdf)

Supplementary Material to "Extreme value methods for estimating rare events in Utopia"

S5.1 Additional figures for Section 5.3

In this section, we present additional figures for Section 5.3, concerning challenges C1 and C2. Figures S5.1-S5.3 support the exploratory analysis for challenges C1 and C2. We explore the within-year seasonality of the response variable Y in Figure S5.1, looking at the distribution of Y per month and across the two seasons. This shows that there is a significant difference in the distribution of Y between seasons 1 and 2, but within each season there is little difference across months.

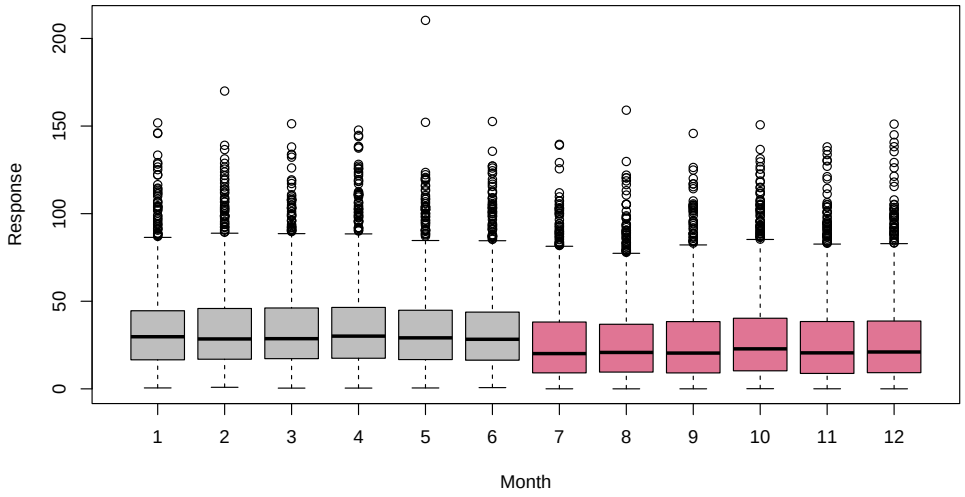

Figure S5.1: Box plot of the response variable Y with each month and season (season 1 in grey and season 2 in red).

Figure S5.2 shows a scatter plot of Y against each covariate $V_1, \ldots, V_8$, excluding V_6 which corresponds to season. Covariates V_1, V_2 and V_8 do not seem to have a relationship with Y, whilst there seems to be dependence for the remaining covariates. These observed relationships appear complex and non-linear.

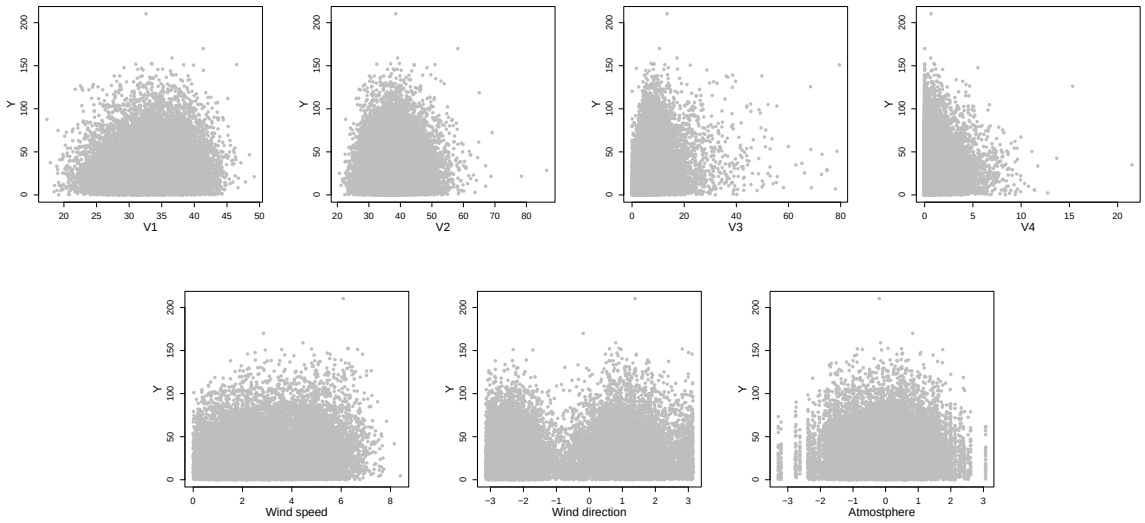

Figure S5.2: Scatter plots of explanatory variables $V_1, \ldots, V_4$, wind speed (V_6), wind direction (V_7) and atmosphere (V_8), from top-left to bottom-right (by row), against the response variable Y.

We also explore temporal dependence in Figure S5.3 that details the auto-correlation function (acf) values for the response Y and explanatory variables $V_1, \ldots, V_4, V_6, \ldots, V_8$, up to a lag of 60. All variables have negligible acf values beyond lag 0, except V_6 (wind speed), V_7 (wind direction) and V_8 (atmosphere).

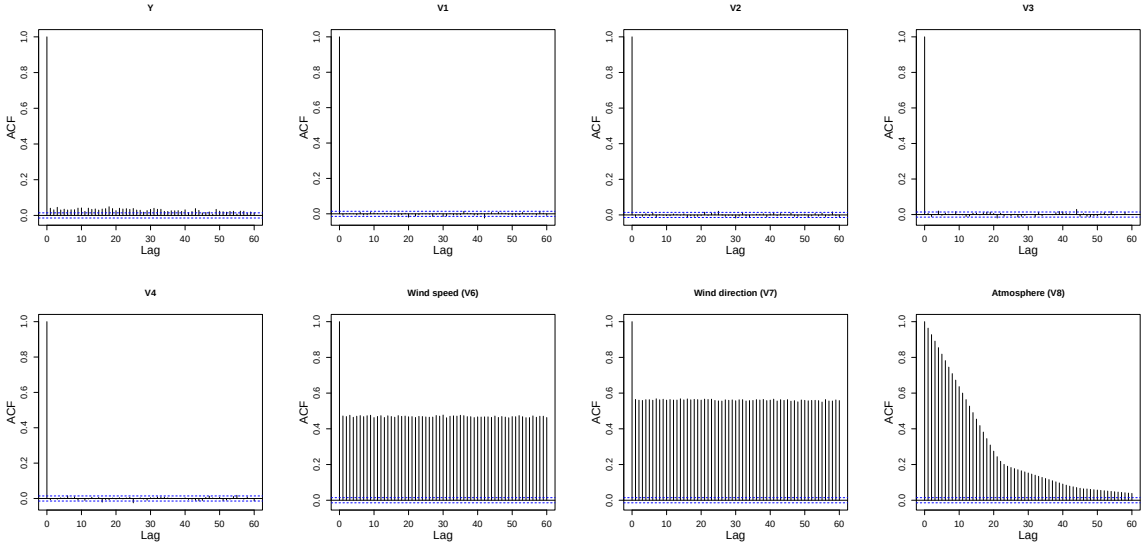

Figure S5.3: Autocorrelation function plots for the response variable Y and explanatory variables $V1, \ldots, V4$, wind speed ($V6$), wind direction ($V7$) and atmosphere ($V8$), from top-left to bottom-right (by row), against the response variable Y.

Figure S5.4 shows the QQ-plots corresponding to a standard GPD model fitted to the excesses

of Y above a constant (left) and seasonally-varying threshold (right). 95% tolerance bounds (grey) show a lack of agreement between observations and the standard GPD model above a constant threshold. The second plot demonstrates a significant improvement in model fit.

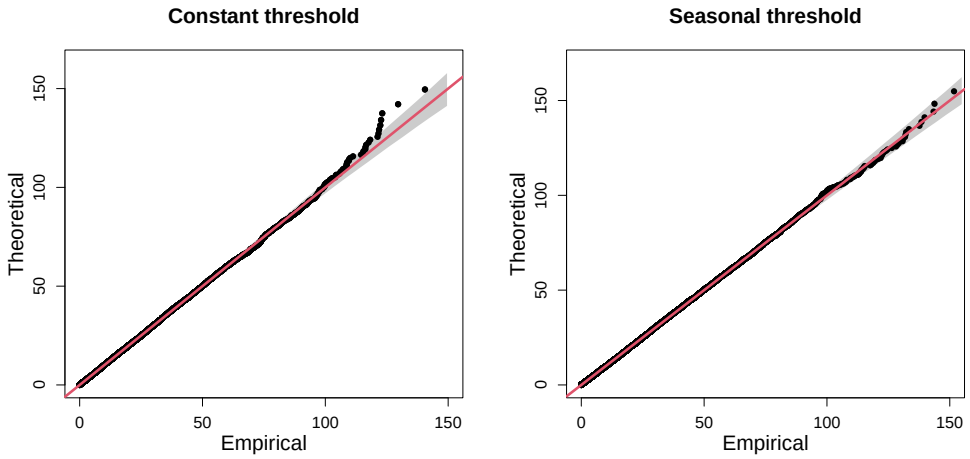

Figure S5.4: QQ-plots showing standard GPD model fits with 95% tolerance bounds (grey) above a constant (left) and stepped-seasonal (right) threshold.

Figure S5.5 shows a detailed summary of the pattern of missing data in the data and can be produced using the `missing_pattern` function in the `finalfit` package in `R` (Harrison et al., 2023). To interpret the figure, note that blue and red squares represent observed and missing variables, respectively. The number on the right indicates the number of missing random variables (i.e., the number of red squares in the row), while the number on the left is the number of observations that fall into the row category. On the bottom, we have the number of observations that fall into the column category. For example, 18,545 observations are fully observed (denoted by the first row); there are 407 observations where only $V4$ is missing (denoted by the second row), 13 observations where both $V4$ and $V6$ are missing (denoted by the fourth row), 456 observations where $V4$ and at least one other predictor is missing (denoted by the last column), etc. It can be seen that there are very few observations where more than one predictor is missing.

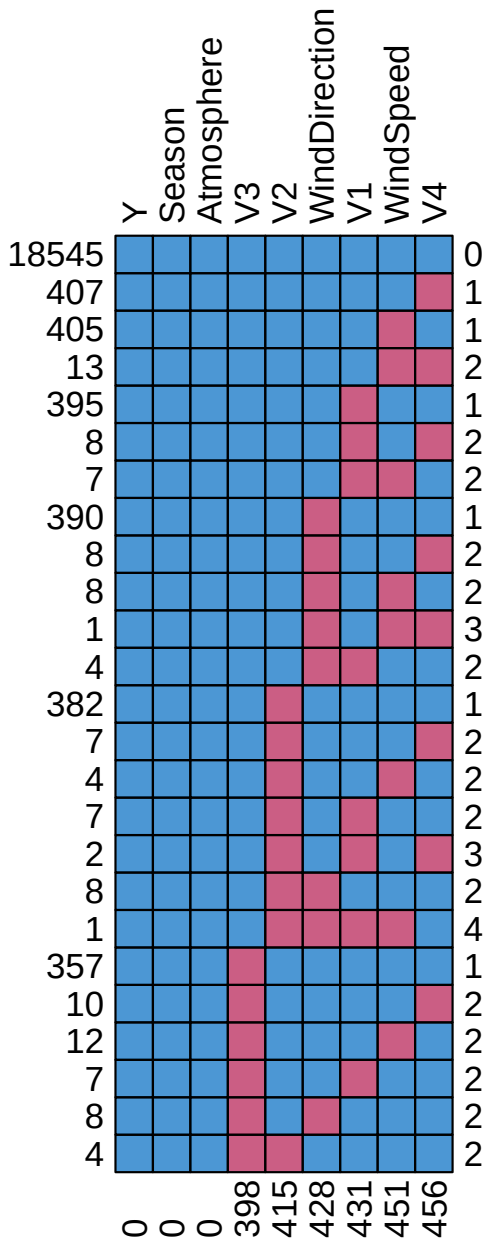

Figure S5.5: Detailed pattern of missing predictor variables in the Amaurot dataset.

S5.2 Additional figures for Section 5.4

In this section, we present additional plots related to Section 5.4. Figure S5.6 illustrates the time series of both covariates for the first 3 years of the observation period. It can be seen how the seasons vary periodically over each year, as well as the discrete nature of the atmospheric covariate.

Bootstrapped χ estimates for the groups $G^A_{I,k}, k \in \{1, \ldots, 10\}$, $I \in \mathcal{I} \setminus \{1, 2, 3\}$ and $G^S_{I,k}, k \in \{1, 2\}$, $I \in \mathcal{I}$ are given in Figures S5.7 - S5.10. These estimates illustrate the impact of

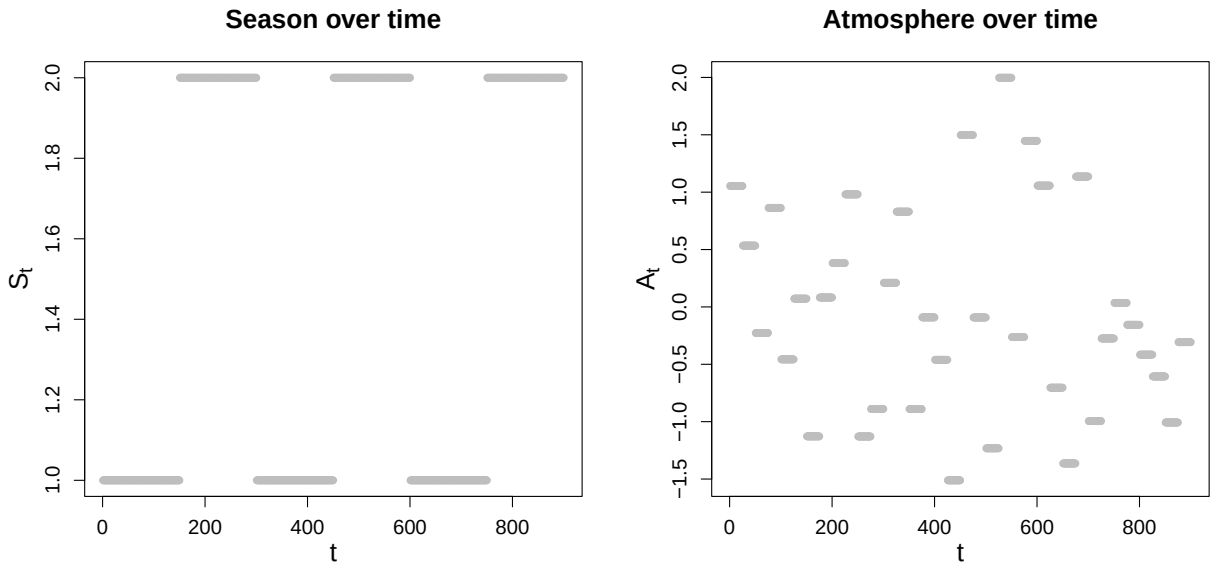

Figure S5.6: Plots of S_t (left) and A_t (right) against t for the first 3 years of the observation period.

atmosphere on the dependence structure.

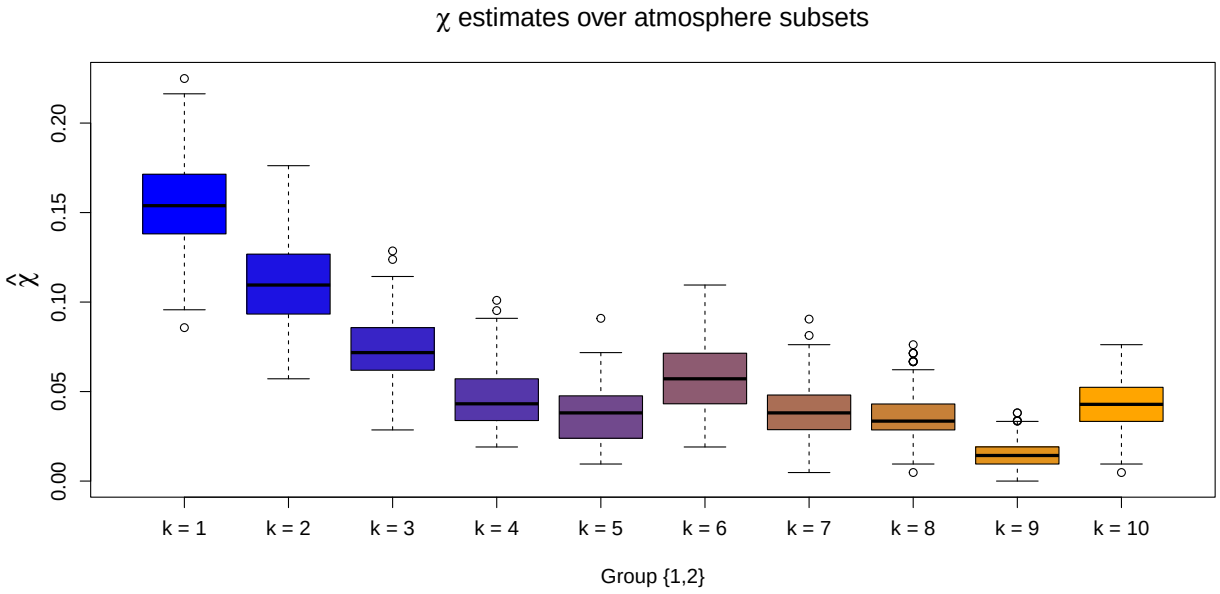

Figure S5.7: Boxplots of empirical χ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

For a 3-dimensional random vector, the angular dependence function, denoted λ, is defined on the unit-simplex $\boldsymbol{S}^2$ and describes extremal dependence along different rays $\boldsymbol{\omega} \in \boldsymbol{S}^2$. As

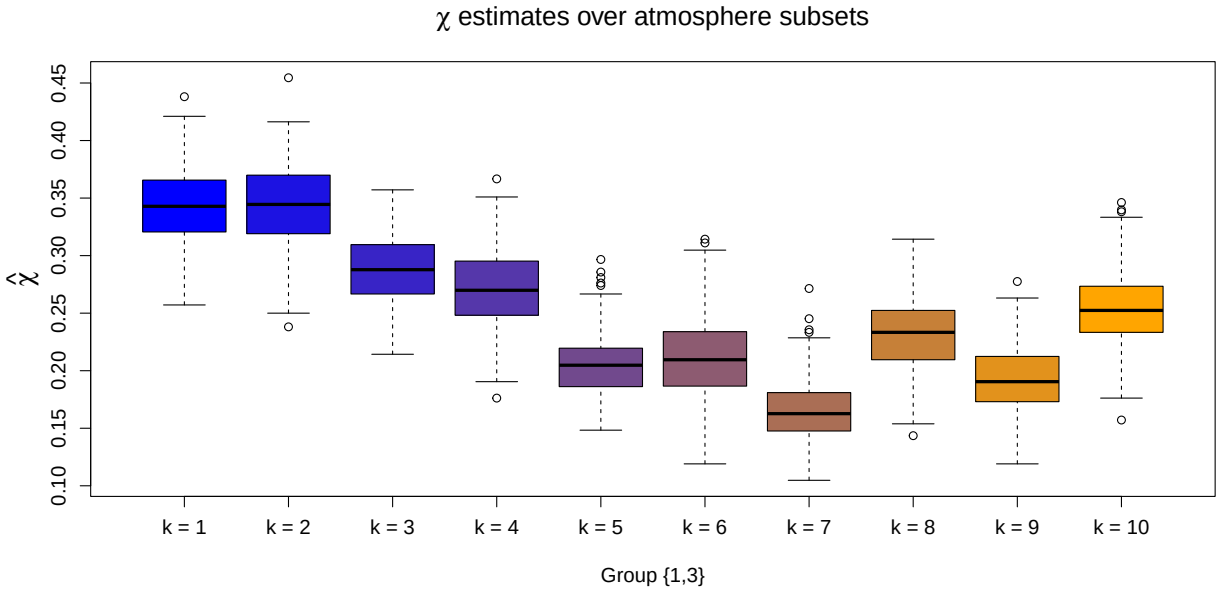

Figure S5.8: Boxplots of empirical χ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

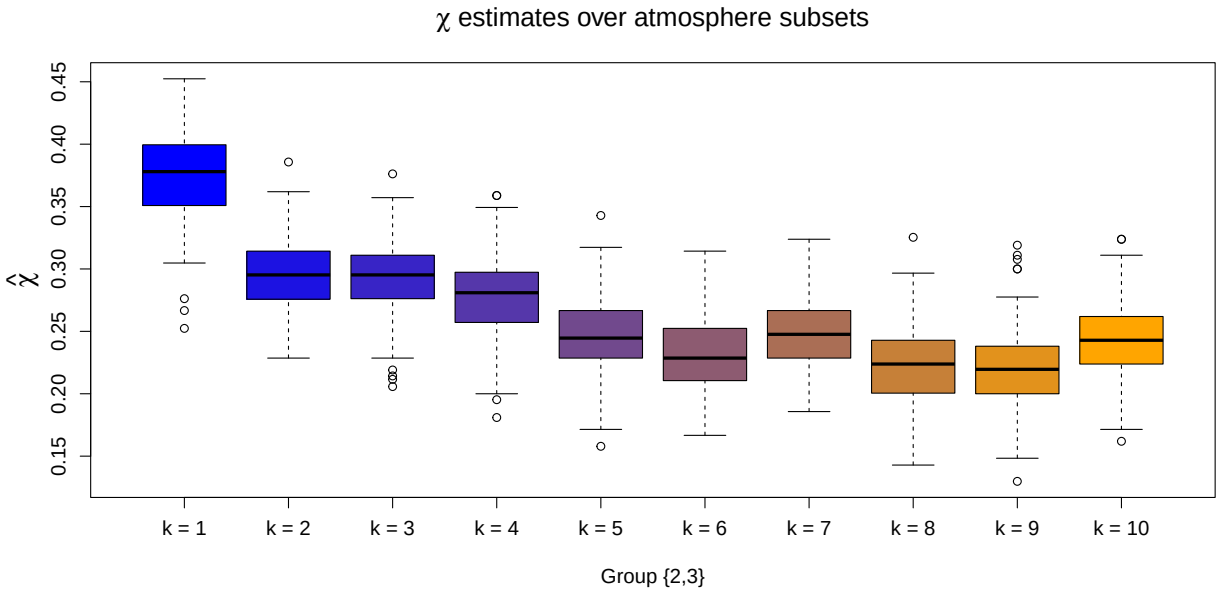

Figure S5.9: Boxplots of empirical χ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

noted in Section 5.4.2, we can associate each of the probabilities from C3, p_1 and p_2, with points on $\boldsymbol{S}^2$, denoted $\boldsymbol{\omega}^1$ and $\boldsymbol{\omega}^2$ respectively. With $I = \{1, 2, 3\}$, we consider $\lambda(\boldsymbol{\omega}^1)$ and $\lambda(\boldsymbol{\omega}^2)$ over the subsets $G^S_{I,k}$, $k \in \{1, 2\}$ and $G^A_{I,k}$, $k \in \{1, \ldots, 10\}$. We note that $\lambda(\boldsymbol{\omega}^1)$

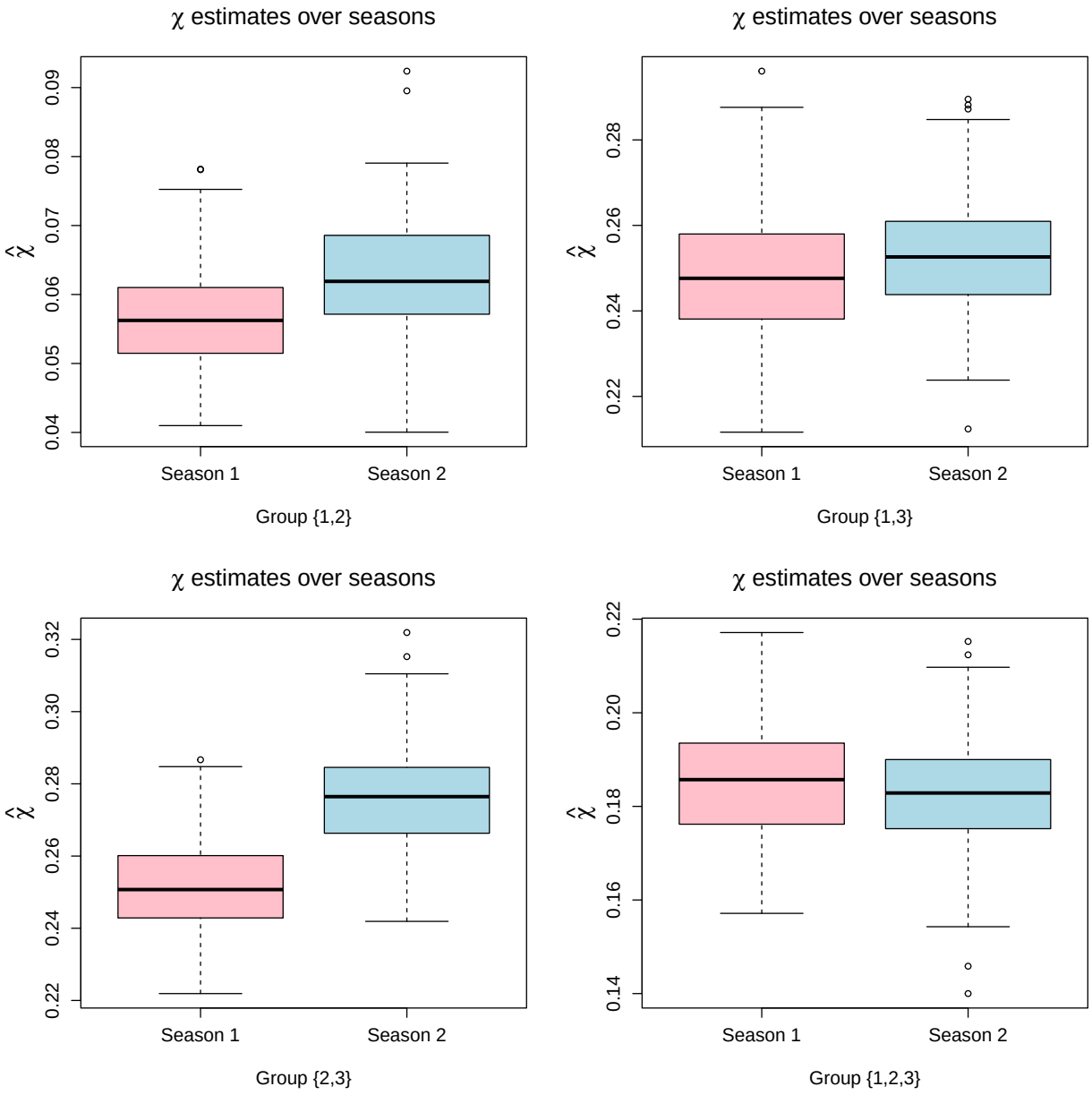

Figure S5.10: Boxplots of empirical χ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively. From top left to bottom right: $I = \{1, 2, 3\}$, $I = \{1, 2\}$, $I = \{1, 3\}$, $I = \{2, 3\}$.

is analogous with the coefficient of tail dependence $\eta \in (0, 1]$ (Ledford and Tawn, 1996), with $\eta = 1/3\lambda(\boldsymbol{\omega}^1)$; this corresponds with the region where all variables are simultaneously extreme. Furthermore, $\lambda(\boldsymbol{\omega}^2)$, which corresponds to a region where only two variables are extreme, is only evaluated after an additional marginal transformation of the third Coputopia time series; see Section 5.4.2.

Estimation of λ for each simplex point and subset was achieved using the Hill estimator

(Hill, 1975) at the 90% level, with uncertainty subsequently quantified via bootstrapping. These results are given in Figures S5.11 - S5.14. These plots provide further evidence of a relationship between the extremal dependence structure and the covariates.

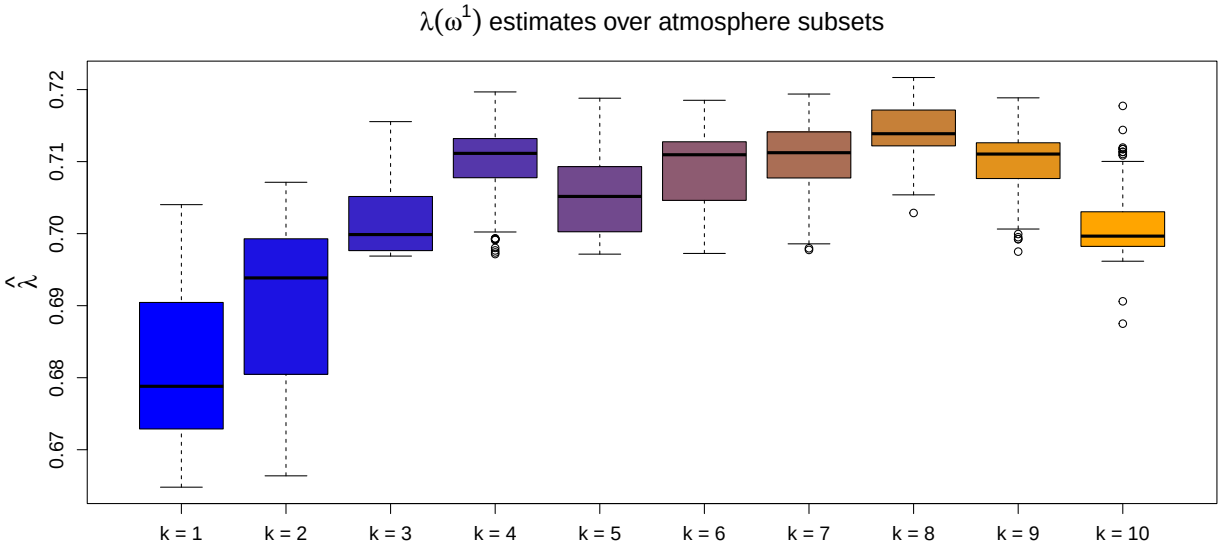

Figure S5.11: Boxplots of empirical $\lambda(\boldsymbol{\omega}_i)$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in λ estimates as the atmospheric values are increased.

To illustrate the estimated trend in dependence, Figure S5.15 shows the estimated scale functions, $\sigma\left(\boldsymbol{\omega} \mid \boldsymbol{x}_t\right)$, over atmosphere for parts 1 and 2. Under the assumption of asymptotic normality in the spline coefficients, 95% confidence intervals are obtained via posterior sampling; see Wood (2017) for more details. We observe that σ tends to increase and decrease over atmosphere for parts 1 and 2, respectively, although the trend is less pronounced for the latter. Under our modelling framework, we note that higher values of σ are associated with less positive extremal dependence in the direction $\boldsymbol{\omega}$ of interest; to see this, observe that the survivor function of the GPD with fixed ξ is negatively associated with σ. Considering the trend in $\sigma\left(\boldsymbol{\omega} \mid \boldsymbol{x}_t\right)$, our results indicate a decrease in dependence in the region where all variables are extreme.

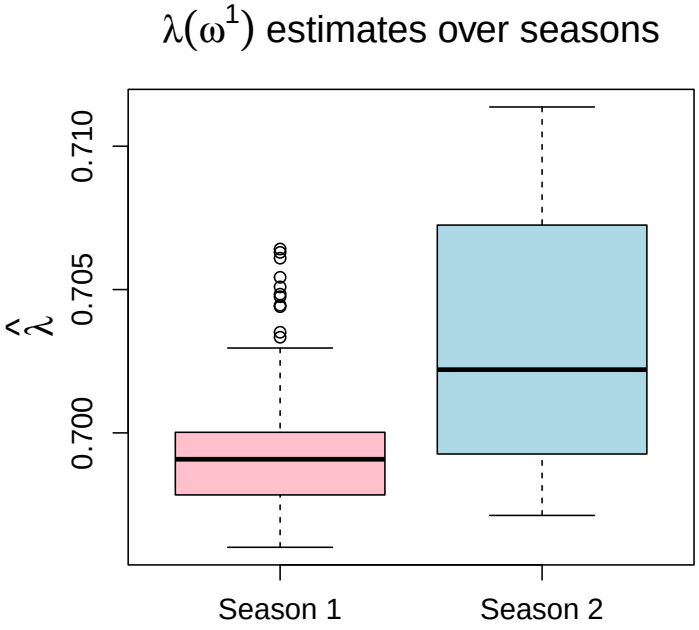

Figure S5.12: Boxplots of empirical $\lambda(\boldsymbol{\omega}_i)$ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$ and $I = \{1, 2, 3\}$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively.

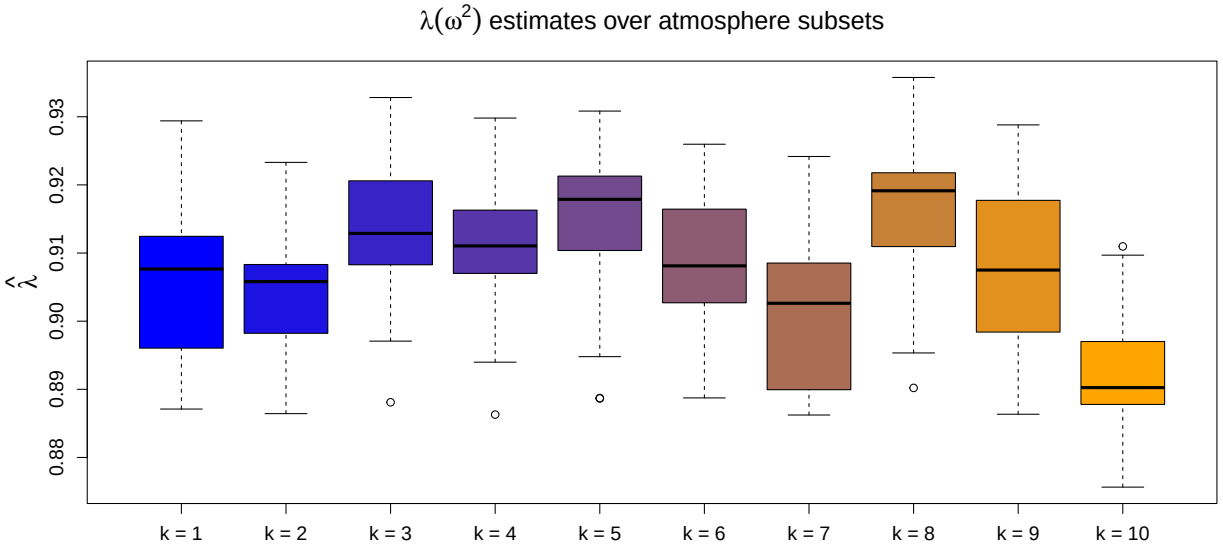

Figure S5.13: Boxplots of empirical $\lambda(\boldsymbol{\omega}_{ii})$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in λ estimates as the atmospheric values are increased.

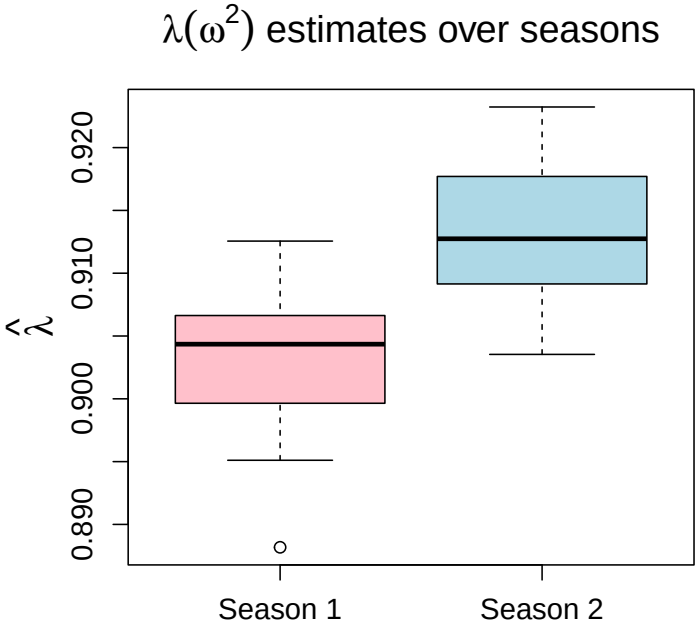

Figure S5.14: Boxplots of empirical $\lambda(\boldsymbol{\omega}_{ii})$ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$ and $I = \{1, 2, 3\}$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively.

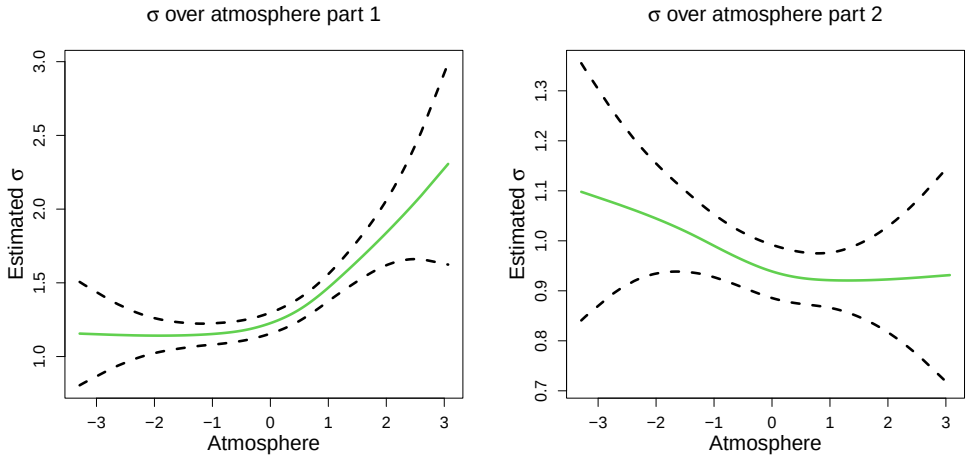

Figure S5.15: Estimated σ functions (green) over atmosphere for part 1 (left) and 2 (right). In both cases, the regions defined by the black dotted lines represent 95% confidence intervals obtained using posterior sampling.

S5.3 Additional figures for Section 5.5

In this section, we present additional plots related to Section 5.5 and we refer to p_1 and p_2 as parts 1 and 2 of C4, respectively. Figure S5.16 shows a heat map of empirically

estimated $\eta(\cdot)$ dependence coefficients and provides further evidence of the existence of the 5 dependence subgroups identified in our exploratory analysis for challenge C4. It also suggests that between group independence as well as within group asymptotic independence – in the sense that the extremes of within group variables do not occur simultaneously – are both reasonable modelling assumptions.

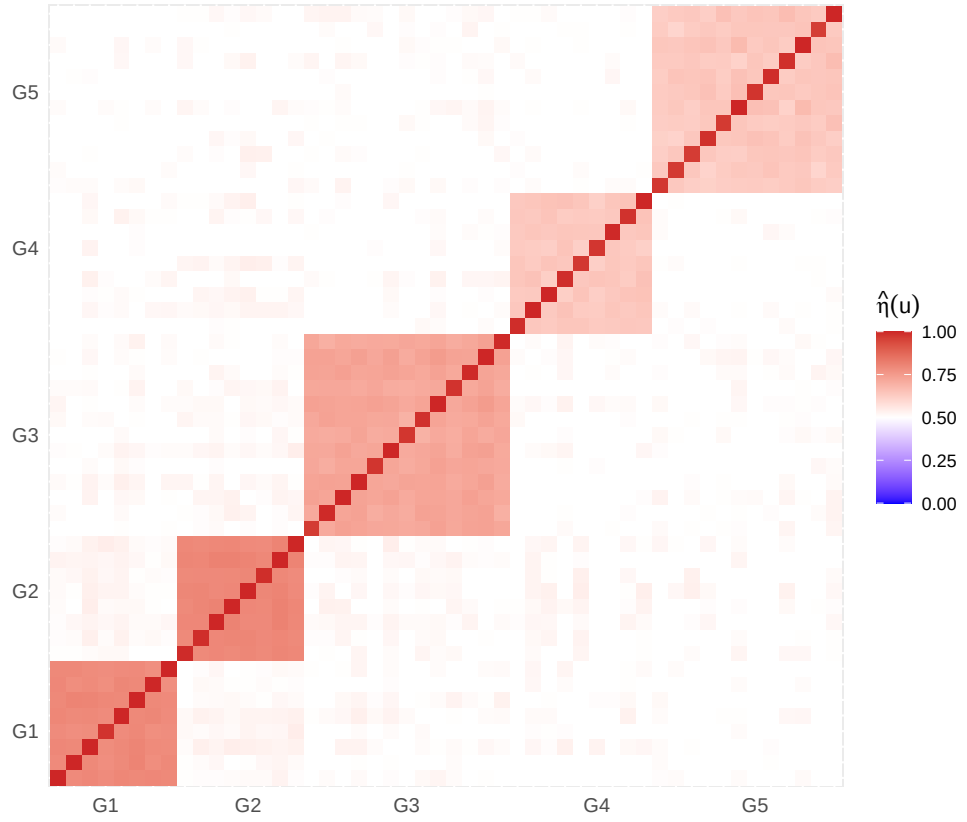

Figure S5.16: Heat map of estimated empirical pairwise $\eta(u)$ extremal dependence coefficients with $u = 0.95$.

Figure S5.17 shows the bootstrapped estimated individual group and overall probabilities with respect to conditioning threshold quantile for part 1 of challenge C4. Similarly, Figure S5.18 shows the bootstrapped estimated individual group and overall probabilities with respect to conditioning threshold quantile for part 2 of challenge C4.

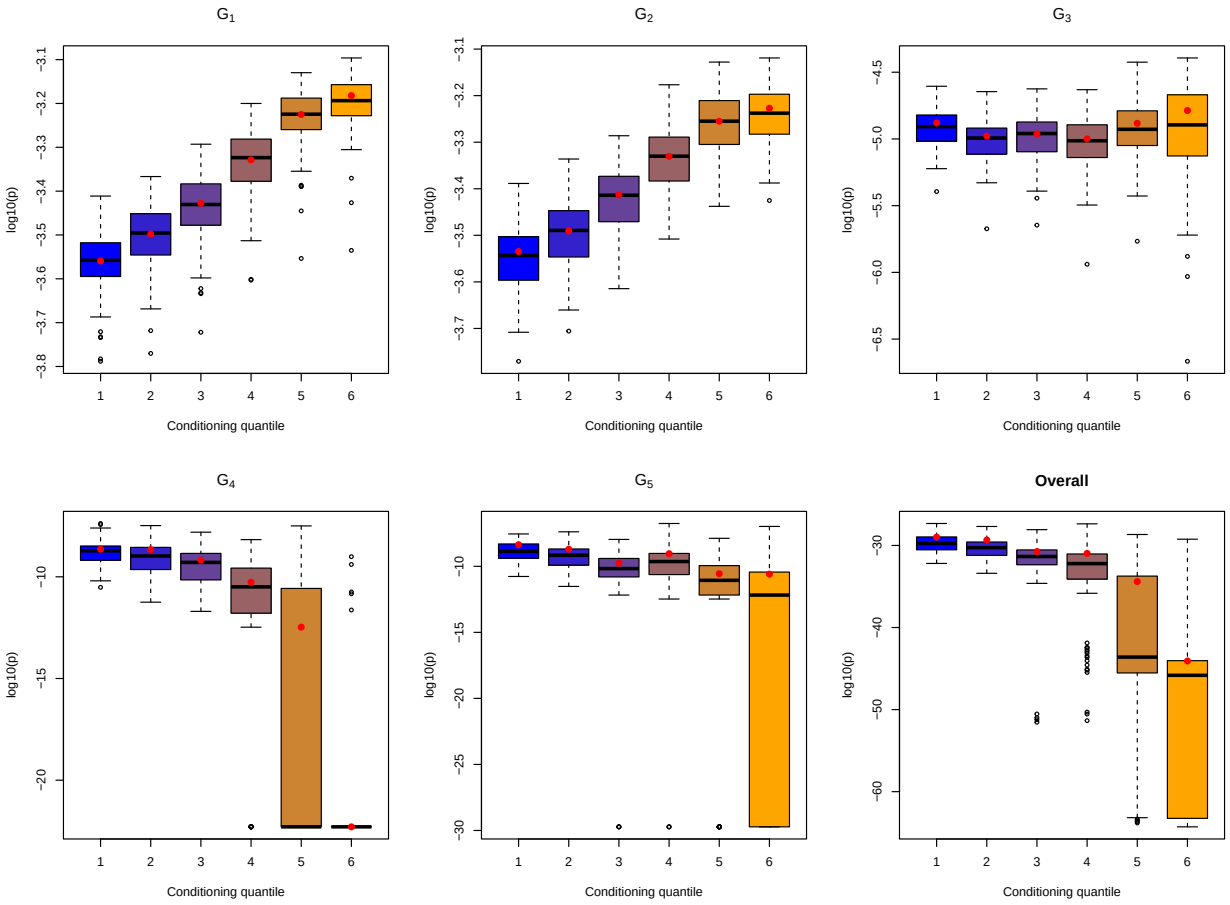

Figure S5.17: Part 1 subgroup and overall bootstrapped probability estimates on the log scale. The red points indicate the original sample estimates and the colouring of the boxplots indicates the choice of conditioning threshold, with the conditioning quantile indices 1-6 referring to the quantile levels $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$, respectively.

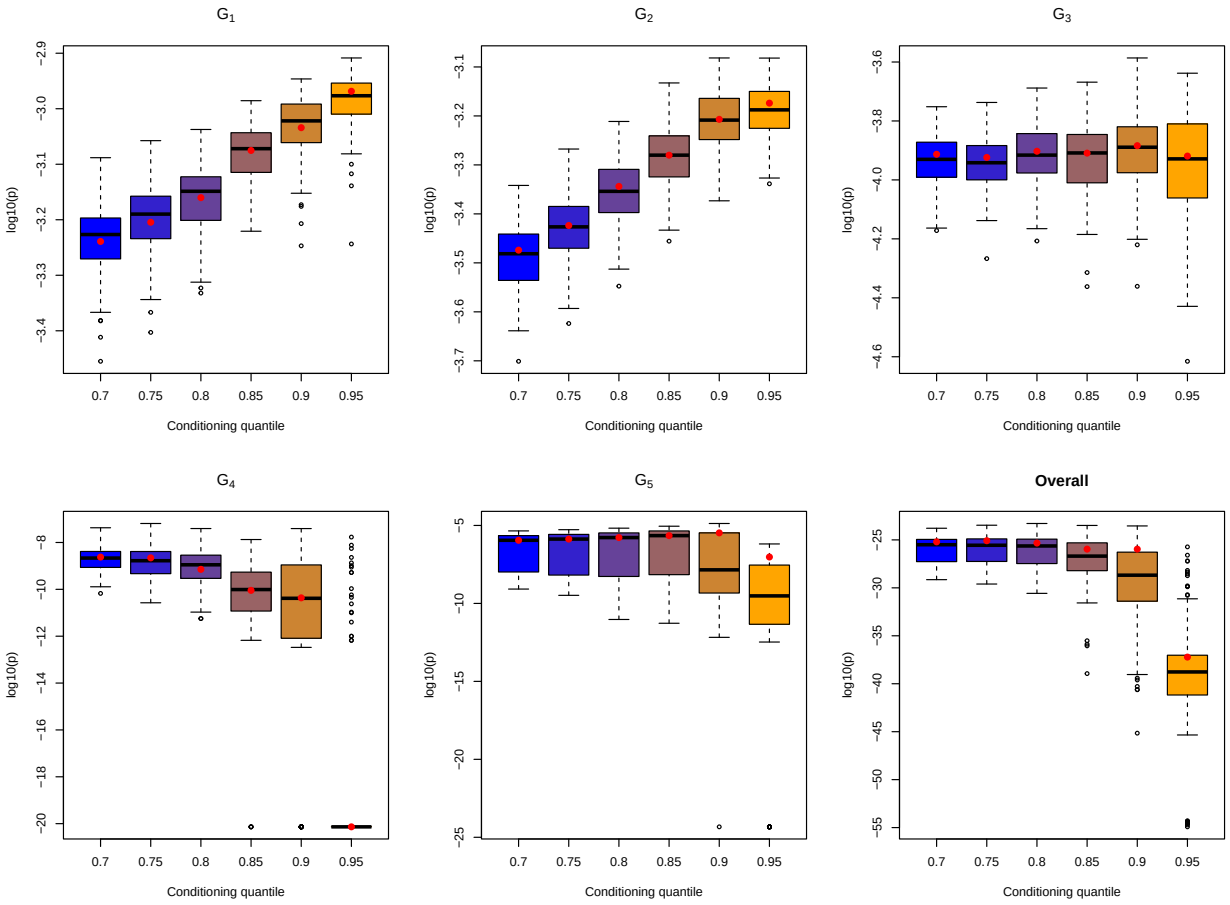

Figure S5.18: Part 2 subgroup and overall bootstrapped probability estimates on the log scale for C4. The red points indicate the original sample estimates and the colouring of the boxplots indicates the choice of conditioning threshold, with the conditioning quantile indices 1-6 referring to the quantile levels $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$, respectively.

Bibliography

Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in france. *Journal of Climate*, 26:7929–7937.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54:207–222.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

D'Arcy, E., Tawn, J. A., Joly, A., and Sifnioti, D. E. (2023). Accounting for seasonality in extreme sea-level estimation. *The Annals of Applied Statistics*, 17(4):3500–3525.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 58:25–45.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Guerrero, M. B., Huser, R., and Ombao, H. (2023). Conex–Connect: Learning patterns in extremal brain connectivity from MultiChannel EEG data. *The Annals of Applied Statistics*, 17:178–198.

Harrison, E., Drake, T., and Ots, R. (2023). *finalfit: Quickly Create Elegant Regression Results Tables and Plots when Modelling*. R package version 1.0.7.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

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*, 66:497–546.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC.

Jonathan, P., Randell, D., Wu, Y., and Ewans, K. (2014). Return level estimation from non-stationary spatial data exhibiting multidimensional covariate effects. *Ocean Engineering*, 88:520–532.

Keef, C., Papastathopoulos, I., and Tawn, J. A. (2013a). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, 115:396–404.

Keef, C., Tawn, J. A., and Lamb, R. (2013b). Estimating the probability of widespread flood events. *Environmetrics*, 24:13–21.

Kyselý, J., Picek, J., and Beranová, R. (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change*, 72:55–68.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Mhalla, L., Opitz, T., and Chavez-Demoulin, V. (2019). Exceedance-based nonlinear regression of tail dependence. *Extremes*, 22:523–552.

Murphy, C., Tawn, J. A., and Varty, Z. (2024). Automated threshold selection and associated inference uncertainty for univariate extremes. *arXiv*, 2310.17999.

Murphy-Barltrop, C. J. R. and Wadsworth, J. L. (2024). Modelling non-stationarity in asymptotically independent extremes. *arXiv*, 2203.05860.

Northrop, P. J. and Jonathan, P. (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*, 22:799–809.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119 – 131.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5:303–336.

Rohrbeck, C., Simpson, E. S., and Tawn, J. A. (2023). Editorial: EVA 2023 Data Challenge. *Extremes*, (to appear).

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107:513–532.

Wadsworth, J. L. and Tawn, J. A. (2013). A new representation for multivariate tail probabilities. *Bernoulli*, 19:2689–2714.

Wood, S. N. (2017). *Generalized Additive Models*. Chapman and Hall/CRC.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *Journal of the American Statistical Association*, 114:1865–1879.

Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software*, 103(3):1–26.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54:437–447.