Modern extreme value statistics for Utopian extremes

EVA (2023) Conference Data Challenge: Team Yalla

Jordan Richards[1*], Noura Alotaibi[2], Daniela Cisneros[2], Yan Gong[3], Matheus B. Guerrero[4], Paolo Redondo[2], Xuanjie Shao[2]

May 2, 2024

Abstract

Capturing the extremal behaviour of data often requires bespoke marginal and dependence models which are grounded in rigorous asymptotic theory, and hence provide reliable extrapolation into the upper tails of the data-generating distribution. We present a toolbox of four methodological frameworks, motivated by modern extreme value theory, that can be used to accurately estimate extreme exceedance probabilities or the corresponding level in either a univariate or multivariate setting. Our frameworks were used to facilitate the winning contribution of Team Yalla to the EVA (2023) Conference Data Challenge, which was organised for the 13$^{\text{th}}$ International Conference on Extreme Value Analysis. This competition comprised seven teams competing across four separate sub-challenges, with each requiring the modelling of data simulated from known, yet highly complex, statistical distributions, and extrapolation far beyond the range of the available samples in order to predict probabilities of extreme events. Data were constructed to be representative of real environmental data, sampled from the fantasy country of "Utopia".

Keywords: additive models; conditional extremes; neural Bayes estimation; non-parametric probability estimators; non-stationary extremal dependence; quantile regression

[1]School of Mathematics, University of Edinburgh, UK
[2]Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.
[3]Harvard School of Public Health, Boston, Massachusetts, USA.
[4]Department of Mathematics, California State University Fullerton, California, USA.
*Corresponding author: jordan.richards@ed.ac.uk

1 Introduction

We are motivated by the EVA (2023) Conference Data Challenge organised for the 13$^{\text{th}}$ International Conference on Extreme Value Analysis, with full details provided in the editorial by Rohrbeck et al. (2023). In this paper, we detail the four modelling frameworks used by "Team Yalla" to win the aforementioned challenge, which comprised four sub-challenges that each required prediction of exceedance probabilities or quantiles for data simulated from highly complex statistical models. These frameworks combined classical EVA methods with modern modelling techniques, including additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), deep learning-based inference (Sainsbury-Dale et al., 2024; Richards et al., 2023b), non-stationary conditional extremal dependence models (Heffernan and Tawn, 2004; Winter et al., 2016), and non-parametric probability estimators (Krupskii and Joe, 2019). Whilst the data considered in this work are simulated, the data were constructed to be representative of real observations of an environmental process (sampled from the fantasy country of "Utopia"), and so exhibit realistic characteristics, such as sparsity, non-stationarity, and missingness. Moreover, as the true values of the challenge predictands are known, our predictions can be easily validated. Hence, we expect our proposed frameworks to perform well in practice, with real data. The code used for implementing our models is available at `https://github.com/matheusguerrero/yalla`.

The novelty of this work is threefold: i) we propose an amortised neural Bayes estimator for univariate quantiles; ii) we generalise the non-parametric multivariate exceedance probability estimator of Krupskii and Joe (2019); and iii) we illustrate the efficacy of the extreme value regression models proposed by Winter et al. (2016) and Youngman (2019) when applied to simulated data. The remainder of the paper is organised as follows: Section 2 describes the methodology we adopt to address the data challenge, with Sections 2.1–2.2 and 2.3–2.5 focusing on univariate and multivariate extremes, respectively. In Section 3, we apply our proposed methodology to the data. Section 4 provides a concise conclusion and suggests avenues for further work.

2 Methodology

In this section, we present the methodological details of our approaches. Throughout, we adopt the notation Y and $\mathbf{Y} := (Y_1, \ldots, Y_d)'$ to denote a random response variable and random d-vector of response variables, respectively. Covariates are denoted by $\mathbf{X} := (X_1, \ldots, X_l)' \in \mathbb{R}^l$ for $l \in \mathbb{N}$. Where appropriate, we use the subscript $t \in \{1, \ldots, n\}$ to denote temporal replicates of the response (i.e., Y_t or $\mathbf{Y}_t := (Y_{1,t}, \ldots, Y_{d,t})'$) or covariates (i.e., $\mathbf{X}_t := (X_{1,t}, \ldots, X_{l,t})'$) and lowercase notation to denote observations. The values of l, d, and sample size n differ throughout the paper. Sub-challenges C1/C2 and C3/C4 (see Rohrbeck et al., 2023) concern univariate and multivariate modelling, respectively. In the univariate setting with $d = 1$ and $n = 21000$, covariate vector $\mathbf{X}$ comprises $l = 8$ variables: wind direction, wind speed, atmosphere, season, and four unnamed variables (denoted $V_1, \ldots, V_4$). For C3 and C4, we have $d = 3$ and $d = 50$, respectively, and $n = 21000$ and $n = 10000$, respectively, with the response vector $\mathbf{Y}$ known to have standard Gumbel margins. The three response variables for C3, (Y_1, Y_2, Y_3), are accompanied by the atmosphere and season covariates described above; hence, $l = 2$ for C3. No covariates accompany the 50 response variables that comprise the data for C4 (i.e., $l = 0$).

Sections 2.1 and 2.2 detail methodology for estimating univariate extreme quantiles, whilst Sections 2.3-2.5 concern methodology for modelling extremal dependence. Section 2.1 describes peaks-over-threshold modelling using the generalised Pareto distribution and generalised additive models, which we used to address sub-challenge C1. Section 2.2 describes a likelihood-free neural Bayes estimator for point estimation of the extreme quantiles for sub-challenge C2. Section 2.3 describes a non-stationary model for conditional extremal dependence (sub-challenge C3), whilst Section 2.4 details a bivariate extremal dependence measure. Section 2.5 concludes with details of a non-parametric estimator for tail probabilities, used in sub-challenge C4.

2.1 Peaks-over-threshold models

Sub-challenge C1 required the prediction of a 50% confidence interval for the $q-$quantile of $Y \mid (\mathbf{X} = \mathbf{x}_j^*)$ for 100 test covariate sets $\{\mathbf{x}_j^* : j = 1, \ldots, 100\}$, and for $q = 0.9999$ corresponding to an extreme conditional quantile. To this end, we adopt a peaks-over-threshold regression model. The upper tails of the distribution of a random variable can be modelled in this framework using the generalised Pareto distribution (GPD), see, for example, Davison and Smith (1990). For a random variable Y, we assume that there exists some high threshold u such that the distribution of exceedances $(Y - u) \mid (Y > u)$ can be characterised by the GPD, denoted GPD(σ_u, ξ), $\sigma_u > 0, \xi \in \mathbb{R}$, with distribution function

$$H(y) = \begin{cases} 1 - (1 + \xi y/\sigma_u)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\sigma_u), & \xi = 0, \end{cases} \tag{1}$$

where $y \geq 0$ for $\xi \geq 0$ and $0 \leq y \leq -\sigma_u/\xi$ for $\xi < 0$.

For regression, we model the conditional distribution $(Y - u(\mathbf{x})) \mid (Y > u(\mathbf{x}), \mathbf{X} = \mathbf{x})$ as GPD$(\sigma_u(\mathbf{x}), \xi(\mathbf{x}))$, where the exceedance threshold and GPD scale and shape parameters are functions of covariates. We follow Chavez-Demoulin and Davison (2005) and Youngman (2019) and use a generalised additive model (GAM) representation for the distribution, with $\sigma_u(\mathbf{x})$ and $\xi(\mathbf{x})$ modelled via a basis of splines. The threshold $u(\mathbf{x})$ is taken to be the intermediate λ-quantile of $Y \mid (\mathbf{X} = \mathbf{x})$ for $\lambda < 0.9999$ and this is modelled using additive quantile regression (Fasiolo et al., 2021).

2.2 Neural point estimation

Sub-challenge C2 required estimation of the q-quantile of Y for $q = 1 - (6 \times 10^4)^{-1}$, i.e., estimating θ such that $\Pr\{Y > \theta\} = 1 - q$. However, inference for θ should seek to minimise the conservative asymmetric loss function

$$L(\theta, \hat{\theta}) = \begin{cases} 0.9(0.99\theta - \hat{\theta}), & \text{if } 0.99\theta > \hat{\theta}, \\ 0, & \text{if } |\theta - \hat{\theta}| \leq 0.01\theta, \\ 0.1(\hat{\theta} - 1.01\theta), & \text{if } 1.01\theta < \hat{\theta}, \end{cases} \tag{2}$$

where $\hat{\theta}$ denotes estimates of θ and $|\cdot|$ denotes the absolute value; this loss function is illustrated in Figure 2. The loss function in Eq. (2) provides a larger penalty for under-estimates of θ, relative to over-estimates, to encourage conservative estimates of the quantile. We construct a conservative estimator for extreme quantiles using neural networks.

Neural point estimators, that is, neural networks that are trained to map input data to parameter point estimates, have shown recent success as a likelihood-free inference approach for statistical models. Although they have been typically used for inference with classical spatial processes (see, e.g., Zammit-Mangion and Wikle, 2020; Gerber and Nychka, 2021) and spatial extremal processes (see, e.g., Lenzi et al., 2023; Lenzi and Rue, 2023; Sainsbury-Dale et al., 2023, 2024; Richards et al., 2023b), they can be exploited in a univariate setting. For example, Rai et al. (2023) use a neural point estimator to make inference with the univariate generalised extreme value distribution (see Coles, 2001). We construct a neural point estimator to perform extreme single quantile estimation for a random variable Z. In particular, we follow Sainsbury-Dale et al. (2024) and construct a neural Bayes estimator (NBE).

Define a set of univariate probability distributions $\mathcal{P}$ on a sample space, taken to be $\mathbb{R}$, which are parameterised by a parameter $\theta \in \mathbb{R}$ such that $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$, where Θ is the parameter space; then $\mathcal{P}$ defines a parametric statistical model (see McCullagh, 2002). Denote $\mathbf{Z} \equiv (Z_1, \ldots, Z_n)'$ as n mutually independent realisations of the random variable Z from $P_\theta \in \mathcal{P}$. A point estimator $\hat{\theta}(\cdot)$ for model $\mathcal{P}$ is any mapping from $\mathbb{R}^n$ to Θ, and the output of such an estimator, for a given θ and $\mathbf{Z}$, can be assessed using a non-negative loss function $L(\theta, \hat{\theta}(\mathbf{Z}))$. The risk of this point estimator, evaluated at θ, $R(\theta, \hat{\theta}(\cdot))$, is the loss $L(\cdot, \cdot)$ averaged over all possible realisations of $\mathbf{Z}$, that is,

$$R(\theta, \hat{\theta}(\cdot)) \equiv \int_{\mathbb{R}^n} L(\theta, \hat{\theta}(\mathbf{z})) f(\mathbf{z} \mid \theta) \, \mathrm{d}\mathbf{z}, \tag{3}$$

where $f(\mathbf{z} \mid \theta)$ is the density function of the data. We define the Bayes risk $r_\pi(\cdot)$ as the

weighted average of Eq. (3) over all $\theta \in \Theta$, with respect to some prior measure $\pi(\cdot)$, as

$$r_\pi(\hat{\theta}(\cdot)) \equiv \int_\Theta R(\theta, \hat{\theta}(\cdot)) \, \mathrm{d}\pi(\theta). \tag{4}$$

If the estimator $\hat{\theta}(\cdot)$ minimises Eq. (4), we term it a *Bayes estimator* with respect to the loss $L(\cdot, \cdot)$ and prior measure $\pi(\cdot)$. Note that in the context of the prediction task, the parameter θ is taken to be the q-quantile of the distribution(s) P_θ and the loss function is $L(\theta, \hat{\theta})$ in Eq. (2) (illustrated in Figure 2). Details on the construction of the prior measure $\pi(\cdot)$ and the statistical model $\mathcal{P}$ will follow.

A neural Bayes estimator (NBE) is a neural network designed to approximately minimise Eq. (4). Sainsbury-Dale et al. (2024) construct NBEs by leveraging the DeepSets neural network architecture (Zaheer et al., 2017). Consider functions $\boldsymbol{\psi} : \mathbb{R} \mapsto \mathbb{R}^Q$ and $\phi : \mathbb{R}^Q \mapsto \mathbb{R}$, and a permutation-invariant set function $\mathbf{a} : (\mathbb{R}^Q)^n \mapsto \mathbb{R}^Q$, where the j-th component of $\mathbf{a}$, $a_j(\cdot)$, returns the element-wise average over its input set for $j = 1, \ldots, Q$. We represent $\phi(\cdot)$ and $\boldsymbol{\psi}(\cdot)$ as neural networks, and collect in $\boldsymbol{\gamma} \equiv (\boldsymbol{\gamma}_\phi', \boldsymbol{\gamma}_{\boldsymbol{\psi}}')'$ their estimable "weights" and "biases". Our NBE is of the form

$$\hat{\theta}(\mathbf{Y}; \boldsymbol{\gamma}) = \phi(\mathbf{T}(\mathbf{Z}; \boldsymbol{\gamma}_{\boldsymbol{\psi}}); \boldsymbol{\gamma}_\phi), \quad \text{with} \quad \mathbf{T}(\mathbf{Z}; \boldsymbol{\gamma}_{\boldsymbol{\psi}}) = \mathbf{a}(\{\boldsymbol{\psi}(Z_t; \boldsymbol{\gamma}_{\boldsymbol{\psi}}) : t = 1, \ldots, n\}). \tag{5}$$

We use a densely-connected neural network to model both $\phi(\cdot)$ and $\boldsymbol{\psi}(\cdot)$. The NBE is built by obtaining neural network weights $\boldsymbol{\gamma}^*$ that minimise the Bayes risk in the estimator space spanned by $\hat{\theta}(\cdot, \boldsymbol{\gamma})$. As we cannot directly evaluate Eq. (4), it is approximated using Monte Carlo methods. For a set of K parameter values $\{\theta^{(k)} : k = 1, \ldots, K\}$ drawn from the prior $\pi(\cdot)$, we simulate, for each k, a set of n mutually independent realisations $\mathbf{z}^{(k)}$ from P_θ. The Bayes risk in Eq. (4) is then approximated by

$$\hat{r}_\pi(\hat{\theta}(\cdot; \boldsymbol{\gamma})) = \frac{1}{K} \sum_{k=1}^{K} L(\theta^{(k)}, \hat{\theta}(\mathbf{z}^{(k)}; \boldsymbol{\gamma})) \approx r_\pi(\hat{\theta}(\cdot; \boldsymbol{\gamma})). \tag{6}$$

We obtain estimates $\boldsymbol{\gamma}^* = \text{argmin}_{\boldsymbol{\gamma}} \hat{r}_\pi(\hat{\theta}(\cdot; \boldsymbol{\gamma}))$ using the package `NeuralEstimators` (Sainsbury-Dale et al., 2024) in `Julia` (Bezanson et al., 2017).

Specifying a prior measure on the q-quantile, $\theta \in \Theta$, for a random variable Z is nontrivial, as

the mapping from the model $\mathcal{P}$ to the parameter space Θ is not guaranteed to be surjective; for a fixed level q, multiple probability distributions can have the same value of θ, and so simulating Z conditional on θ is not necessarily feasible. In this case, P_θ would define a set of distributions with equal q-quantile, rather than a single probability distribution (as in Sainsbury-Dale et al. (2024)). Thus, instead of placing a prior directly on θ, we assume that P_θ is determined by some hyper-parameters (for simplicity, we omit the dependency of P_θ on hyper-parameters from our notation). We then construct a general prior measure for these hyper-parameters, which consequently induces a prior measure on θ. As we are interested in θ when q is close to one, our choice of the class of feasible models $\mathcal{P}$ is motivated by extreme value theory, and we exploit the univariate peaks-over-threshold models described in Section 2.1.

Rohrbeck et al. (2023) note that our data $\{Y_t : t = 1, \dots, n\}$, from which we wish to infer θ, are non-stationary over time. We reflect this property in our construction for the prior measure on the hyper-parameters of P_θ: for $t = 1, \dots, n$, let $(Y_t - u_t) \mid (Y_t > u_t) \sim \text{GPD}(\sigma_t, \xi_t)$ and let $Y_t \mid (Y_t \leq u_t) \sim F_t^{\leq}(\cdot)$ for threshold $u_t \in \mathbb{R}$, scale $\sigma_t > 0$, and shape $\xi_t \in \mathbb{R}$, and where $\Pr\{Y_t \leq u_t\} = \lambda$ for $\lambda \in [0, 1]$. For simplicity, we hereafter treat λ and the distribution of non-exceedances $F_t^{\leq}(\cdot)$ as fixed. After placing a suitable prior measure on the hyper-parameters $\{(u_t, \sigma_t, \xi_t) : t = 1, \dots, n\}$ and specifying λ and $F_t^{\leq}(\cdot)$ (see, e.g., Section 3.1), we can simulate data $\{y_t^* : t = 1, \dots, n^*\}$ from the model above and store this in a vector $\mathbf{z}^*$ with the index t removed from each entry. In this way, we consider $\mathbf{z}^*$ as n^* independent draws from the distribution of Z, unconditional on t, as we have marginalised out the effect of this covariate (see, e.g., Rohrbeck et al., 2018). Note that n^* need not satisfy $n^* = n$, where n is the sample size.

The prior measure on the hyper-parameter set $\{(u_t, \sigma_t, \xi_t) : t = 1, \dots, n\}$ induces a prior on θ, which is the q-quantile of Z. As there is no closed-form expression for θ, we compute it using Monte Carlo methods. That is, we set n^* large and derive θ empirically from realisations $\mathbf{z}^*$. A single entry to the training data for our NBE then consists of the

pair $(\mathbf{z}, \theta)$, where $\mathbf{z}$ is a sub-sample from $\mathbf{z}^*$ of length n; this procedure is repeated for a total of K entries. Note that, we could instead construct a prior on θ using a stationary GPD model, i.e., with $n = 1$, but our approach produces a more diffuse prior on θ by increasing the prior support for the hyper-parameters of P_θ. It also exploits knowledge about the data generating distribution, which produces more realistic models for Z.

When choosing larger values of λ, fewer new observations are generated for training of the NBE; a bigger proportion of the training data comprise resamples from the observations. Smaller values of λ will produce a more diffuse prior on the distributional models for Z (particularly with regards to the upper-tail behaviour of Z) and, hence, the q-quantile of Z. A larger variety in the training data is also likely to increase the reliability of our estimator when generalising to unseen data. However, when λ is too small, the threshold u_t may be too low to safely assume that $(Y_t - u_t) \mid (Y_t > u_t)$ follows a GPD. In this case, our estimator would not benefit from being trained on well-specified models for Y_t. Hence, we advocate taking λ as low as possible whilst still obtaining reasonable fits for the non-stationary GPD model, even if this value is not optimal. In our application, we take $\lambda = 0.6$ but note that this is lower than the optimal λ (in terms of providing the best fit for the model described in Section 2.1).

2.3 Conditional extremes models

Sub-challenge C3 required estimation of

$$p_1 := \Pr(Y_1 > 6, Y_2 > 6, Y_3 > 6), \tag{7}$$
$$p_2 := \Pr(Y_1 > 7, Y_2 > 7, Y_3 < -\log(\log 2)),$$

where $-\log(\log 2)$ is the median of the standard Gumbel distribution. To estimate these extreme exceedance probabilities, we construct a non-stationary extremal dependence model for random vectors. Proposed by Heffernan and Tawn (2004) and later generalised by Heffernan and Resnick (2007), the conditional extremes framework models the behaviour of a random vector, conditional on one of its components being extreme. We adopt this model

as its inference is typically less computationally demanding than that of other models for multivariate extremal dependence (Huser and Wadsworth, 2022). Additionally, it is capable of capturing both asymptotic dependence and asymptotic independence in a parsimonious manner. To accommodate non-stationarity with respect to covariates in the extremal dependence structure of our data, we utilise the extension of the Heffernan and Tawn (2004) model proposed by Winter et al. (2016). This extension represents the dependence parameters as a linear function (subject to some non-linear link transformation) of the covariates.

Let the vector $\mathbf{Y}_t \equiv (Y_{1,t}, Y_{2,t}, Y_{3,t})'$ for $t \in \{1, \ldots, n\}$ have standard Laplace margins (Keef et al., 2013). It is noteworthy that our original data are known to possess standard Gumbel marginals; therefore, we transform these to have standard Laplace margins. Then, denote by $\mathbf{Y}_{-i,t}$ as the vector $\mathbf{Y}_t$ with its i-th component removed. Note that all vector operations hereafter are taken component-wise. Winter et al. (2016) assume that there exist vectors of coefficients $\boldsymbol{\alpha}_{-i,t} := \{\alpha_{j|i,t} : j \in (1,2,3) \setminus i\} \in [-1,1]^2$ and $\boldsymbol{\beta}_{-i,t} := \{\beta_{j|i,t} : j \in (1,2,3) \setminus i\} \in [0,1)^2$ such that, for $\mathbf{z} \in \mathbb{R}^2$ and $y > 0$,

$$\Pr\left\{ \frac{\mathbf{Y}_{-i,t} - \boldsymbol{\alpha}_{-i,t} Y_{i,t}}{Y_{i,t}^{\boldsymbol{\beta}_{-i,t}}} \leq \mathbf{z}, Y_{i,t} - u > y \;\middle|\; Y_{i,t} > u \right\} \to G_{-i,t}(\mathbf{z}) \exp(-y), \tag{8}$$

as $u \to \infty$ with $G_{-i,t}(\cdot)$ a non-degenerate bivariate distribution function. The values of the dependence parameters $\boldsymbol{\alpha}_{-i,t}$ and $\boldsymbol{\beta}_{-i,t}$ determine the strength and class of extremal dependence exhibited between $Y_{i,t}$ and the corresponding component of $\mathbf{Y}_{-i,t}$; for details, see Heffernan and Tawn (2004). We allow these parameters to vary with covariates $\mathbf{x}_t$ by letting

$$\tanh^{-1}(\boldsymbol{\alpha}_{-i,t}) := \boldsymbol{\alpha}_{-i}^{(0)} + \boldsymbol{\alpha}_{-i}^{(1)} \mathbf{x}_t, \quad \text{logit}(\boldsymbol{\beta}_{-i,t}) := \boldsymbol{\beta}_{-i}^{(0)} + \boldsymbol{\beta}_{-i}^{(1)} \mathbf{x}_t, \tag{9}$$

with coefficients $\boldsymbol{\alpha}_{-i}^{(0)}, \boldsymbol{\beta}_{-i}^{(0)} \in \mathbb{R}^2$ and $\boldsymbol{\alpha}_{-i}^{(1)}, \boldsymbol{\beta}_{-i}^{(1)} \in \mathbb{R}^{2 \times l}$, where l is the number of covariates. Note that the $\tanh(\cdot)$ and $\text{logit}(\cdot)$ link functions are used to ensure that the parameter values are constrained to their correct ranges.

Modelling follows under the assumption that the limit in Eq. (8) holds in equality for all $Y_{i,t} > u$ for some sufficiently high threshold $u > 0$. In this case, rearranging Eq. (8) provides

the model

$$\mathbf{Y}_{-i,t} = \boldsymbol{\alpha}_{-i,t} Y_{i,t} + Y_{i,t}^{\boldsymbol{\beta}_{-i,t}} \mathbf{Z}_{i,t} \mid (Y_{i,t} > u), \tag{10}$$

where the residual random vector $\mathbf{Z}_{i,t} := \{Z_{j|i,t} : j \in (1,2,3) \setminus i\} \sim G_{-i,t}$ is independent of $Y_{i,t}$. For inference, we make the working assumption that $G_{-i,t}(\cdot)$ does not depend on time t and follows a bivariate standard Gaussian copula with correlation $\rho_i \in (-1,1)$ and delta-Laplace margins (see, e.g., Shooter et al., 2021). We hereafter drop the subscript t from the notation. A random variable that follows the delta-Laplace distribution with location, scale, and shape parameters $\mu \in \mathbb{R}$, $\sigma > 0$, and $\delta > 0$, respectively, has density function $f(z) = \delta(2k\sigma\Gamma(\delta^{-1}))^{-1} \exp\{-(|z-\mu|/(k\sigma))^{\delta}\}$ for $k^2 = \Gamma(\delta^{-1})/\Gamma(3\delta^{-1})$, where $\Gamma(\cdot)$ denotes the standard gamma function. Note that when $\delta = 1$ and $\delta = 2$, we have the Laplace and Gaussian densities, respectively.

Inference proceeds via maximum likelihood estimation. The model is fitted separately for each conditioning variable $i = 1, 2, 3$. For each i, we have eight parameters in $\boldsymbol{\alpha}_{-i}^{(0)}$, $\boldsymbol{\alpha}_{-i}^{(1)}$, $\boldsymbol{\beta}_{-i}^{(0)}$, $\boldsymbol{\beta}_{-i}^{(1)}$, as well as the seven parameters that characterise G_{-i}, that is, the correlation ρ_i and the three marginal parameters for each component of $\mathbf{Z}_{i,t}$. After estimation of the parameters, we no longer require the working Gaussian copula assumption for $\mathbf{Z}_{i,t}$. We instead use observations $\mathbf{y}_{-i,t}$ and $y_{i,t}$ to derive the empirical residual vector

$$\tilde{\mathbf{z}}_{i,t} := (\mathbf{y}_{-i,t} - \hat{\boldsymbol{\alpha}}_{-i,t} y_{i,t}) / {y_{i,t}}^{\hat{\boldsymbol{\beta}}_{-i,t}},$$

where $\hat{\boldsymbol{\alpha}}_{-i,t}$ and $\hat{\boldsymbol{\beta}}_{-i,t}$ denote estimates of $\boldsymbol{\alpha}_{-i,t}$ and $\boldsymbol{\beta}_{-i,t}$, respectively. Assuming independence across time, we use the empirical residuals to provide an empirical estimate $\tilde{G}_{-i}(\cdot)$ of $G_{-i}(\cdot)$.

We estimate the necessary exceedance probabilities, p_1 and p_2 in Eq. (7), via the following Monte-Carlo procedure. We first note that, for fixed $t \in \{1, \ldots, n\}$, we require realisations of $\mathbf{Y}_t$, i.e., unconditional on an exceedance. We follow, for example, Richards et al. (2022, 2023c) and obtain these realisations by drawing a realisation from

$$\mathbf{Y}_t \;\Big|\; \left(\max_{i=1,2,3} Y_{i,t} > u \right), \tag{11}$$

with probability

$$\Pr\left\{\max_{i=1,2,3} Y_{i,t} > u\right\}, \tag{12}$$

and, otherwise, drawing a realisation of $\mathbf{Y}_t \mid (\max_{i=1,2,3} Y_{i,t} < u)$. As realisations of the latter are unlikely to significantly impact estimates of p_1 and p_2, we draw them empirically (Richards et al., 2022). In practice, we also replace the probability in Eq. (12) with an empirical estimate. In addition, as both $\mathbf{Y}_t \mid (\max_{i=1,2,3} Y_{i,t} < u)$ and Eq. (12) depend on t, we estimate them empirically by assuming stationarity over time.

To simulate from Eq. (11), we must first draw realisations of the conditional exceedance model, $\mathbf{Y}_{-i,t} \mid (Y_{i,t} > u)$ in Eq. (10), using Algorithm 1. We can then combine realisations

Algorithm 1 Simulating from Eq. (10).

For time t and conditioning index $i \in \{1, 2, 3\}$:

1. Simulate $E \sim \text{Exp}(1)$ and set $y_{i,t} = u + E$.
2. Draw a residual vector $\tilde{\mathbf{z}} \sim \tilde{G}_{-i}(\cdot)$.
3. Set $\mathbf{y}_{-i,t} = \hat{\boldsymbol{\alpha}}_{-i,t} y_{i,t} + y_{i,t}^{\hat{\boldsymbol{\beta}}_{-i,t}} \tilde{\mathbf{z}}$.

from the three separate conditional exceedance models, i.e., $\mathbf{Y}_{-i,t} \mid (Y_{i,t} > u)$ for each $i = 1, 2, 3$, into a single realisation of Eq. (11) using importance sampling (Wadsworth and Tawn, 2022). Whilst this can be achieved for a single fixed value of t, we note that p_1 and p_2 in Eq. (7) do not depend on the time t. Hence, we treat t as random during simulation and average over all times $t \in \{1, \ldots, n\}$ to produce approximate realisations of $\mathbf{Y}$ with t marginalised out. The full simulation algorithm is detailed in Algorithm 2. Note that we back-transform from standard Laplace to standard Gumbel margins after simulation, as the original data has the latter.

2.4 Extremal dependence measure

A number of pairwise measures have been proposed to quantify the strength of extremal dependence between random variables (Y_1, Y_2), see, for example, Heffernan (2000). We

Algorithm 2 Simulating $\{(Y_{1,t}, Y_{2,t}, Y_{3,t}) : t \in \{1, \ldots, n\}\}$

1. For $j = 1, \ldots, N'$ with $N' > N$:
   (a) Draw a conditioning index $i_j \in \{1, 2, 3\}$ with equal probability.
   (b) Draw a time $t_j \in \{1, \ldots, n\}$ with equal probability.
   (c) Simulate $\mathbf{y}^{(j)}$ using Algorithm 1 with time t_j and conditioning index i_j.
2. Assign each simulated vector $\mathbf{y}^{(j)} := (y_1^{(j)}, y_2^{(j)}, y_3^{(j)})'$ an importance weight of
$$\left\{ \sum_{i=1}^{3} \mathbb{1}\{y_i^{(j)} > u\} \right\}^{-1},$$
for $j = 1, \ldots, N'$, and sub-sample N realisations from the collection with probabilities proportional to these weights.
3. With probability $n^{-1} \sum_{t=1}^{n} \mathbb{1}\{\max_{i=1,2,3} y_{i,t} < u\}$, re-sample (with equal probability) from
$$\left\{ \mathbf{y}_t : t \in \{1, \ldots, n\}, \max_{i=1,2,3} y_{i,t} < u \right\}.$$
4. Back-transform sample onto standard Gumbel margins.

adopt the *extremal dependence measure* (EDM) proposed by Resnick (2004) and Larsson and Resnick (2012), which has been extended to a multivariate setting by Cooley and Thibaud (2019). Let $(Y_1, Y_2) \in [0, \infty)^2$ be a regularly-varying random vector with index $\alpha > 0$; for full details on regular variation, see Resnick (2007). For some symmetric norm $\|\cdot\|$ on $[0, \infty)^2$, define the transformation $(R, \boldsymbol{\Omega}) := (\|(Y_1, Y_2)\|, (Y_1, Y_2)/\|(Y_1, Y_2)\|)$. Then, there exists a sequence $b_n \to \infty$ and constant $c > 0$ such that

$$n \Pr\{(b_n^{-1} R, \boldsymbol{\Omega}) \in \cdot\) \xrightarrow{v} c\nu_\alpha \times H,$$

where $\xrightarrow{v}$ denotes vague convergence, ν_α is a measure on $(0, \infty]$ such that $\nu_\alpha((y, \infty]) = y^{-\alpha}$ and the angular measure H is defined on the unit circle $\mathcal{S}_+ = \{\mathbf{y} \in [0, \infty)^2 \setminus \{\mathbf{0}\} : \|\mathbf{y}\| = 1\}$. The extremal dependence measure (Larsson and Resnick, 2012) between Y_1 and Y_2 is

$$\mathrm{EDM}(Y_1, Y_2) = \int_{\mathcal{S}_+} \omega_1 \omega_2 \, \mathrm{d}H(\boldsymbol{\omega}) = \lim_{y \to \infty} \mathbb{E}\left(\Omega_1 \Omega_2 \mid R > y\right).$$

Note that $\text{EDM}(Y_1, Y_2) \in [0, 1]$ quantifies the strength of extreme dependence between Y_1 and Y_2. If $\text{EDM}(Y_1, Y_2) = 0$, then H concentrates all mass along the axes and hence Y_1 and Y_2 are asymptotically independent (Resnick, 2007). Conversely, $\text{EDM}(Y_1, Y_2)$ is maximal if and only if (Y_1, Y_2) has full asymptotic dependence, or H places all mass on the diagonal. Larsson and Resnick (2012) propose the empirical estimator

$$\widehat{\text{EDM}}\,(Y_1, Y_2) = \frac{1}{N_n} \sum_{t=1}^{n} \frac{y_{1,t}}{r_t} \frac{y_{2,t}}{r_t} \mathbb{1}\{r_t > u\}, \tag{13}$$

where $\boldsymbol{y}_t := (y_{1,t}, y_{2,t})$, $t = 1, \ldots, n$, is an i.i.d sample from (Y_1, Y_2), $r_t = \|\boldsymbol{y}_t\|$, and $N_n = \sum_{t=1}^{n} \mathbb{1}\{r_t > u\}$ is the number of exceedances of $\{r_t : t = 1, \ldots, n\}$ above some high threshold $u > 0$. In Section 3.2, we use pairwise EDM estimates to investigate the extremal dependence structure of a high-dimensional random vector and then decompose it into subvectors of strongly tail-dependent variables.

2.5 Non-parametric tail probability estimation

Sub-challenge C4 requires estimation of two exceedance probabilities of the form

$$\begin{aligned} p_3 &:= \Pr(Y_1 > s_1, \ldots, Y_{50} > s_1), \\ p_4 &:= \Pr(Y_1 > s_1, \ldots, Y_{25} > s_1, Y_{26} > s_2, \ldots, Y_{50} > s_2), \end{aligned} \tag{14}$$

where $s_j, j = 1, 2$, is the $(1 - \phi_j)$-quantile of the standard Gumbel distribution with $\phi_1 = 1/300$ and $\phi_2 = 12\phi_1$. Hence, p_3 corresponds to an exceedance probability with all components of $\mathbf{Y}$ being equally extreme, i.e., concurrently exceeding the same marginal quantile, whilst, for p_4, the first 25 components of $\mathbf{Y}$ exceed a higher quantile, i.e., are more extreme, than the latter 25 components. Thus, we seek an estimator for exceedance probabilities of high-dimensional multivariate random vectors $\mathbf{Y} \in \mathbb{R}^d$. While scalable parametric models for high-dimensional multivariate extremes do exist, as exemplified by Engelke and Ivanovs (2021) and Lederer and Oesting (2023), they often impose restrictive assumptions about extremal dependence in $\mathbf{Y}$, particularly as the dimension d grows large. Examples

include regular or hidden regular variation (Resnick, 2002). As our goal is point estimation of exceedance probabilities, rather than a full characterisation of the joint upper tail of $\mathbf{Y}$, we choose instead to use a non-parametric estimator for exceedance probabilities that makes few assumptions on the joint upper tails. In particular, we adopt the non-parametric multivariate tail probability estimator proposed by Krupskii and Joe (2019).

Denote by $F_i(\cdot)$ the distribution function[1] of Y_i and let $U_i = F_i(Y_i)$ for $i = 1, \ldots, d$. Then, for $\phi \in (0, 1)$, let

$$p(\phi) := \Pr\{U_1 > 1 - \phi, \ldots, U_d > 1 - \phi\} = \Pr\{U_{\max} < \phi\}, \tag{15}$$

where $U_{\max} = \max(U_1, \ldots, U_d)$. Denote by $C : [0, 1]^d \mapsto \mathbb{R}$ the copula associated with $\mathbf{Y}$ and by $\bar{C}$ the corresponding survival copula, such that $p(\phi) = \bar{C}(1 - \phi, \ldots, 1 - \phi)$. Krupskii and Joe (2019) construct estimators of $p(\phi)$ for small $\phi > 0$ by making assumptions about the joint tail decay of C. In particular, they assume that C (and $\bar{C}$) has continuous partial derivatives and, as $\phi \downarrow 0$, that

$$\begin{aligned} p(\phi) &= \Pr\{U_{\max} < \phi\} = \lambda_1 \phi + \lambda_2 \phi^{1/\eta} \ell(\phi) + o\{\phi^{1/\eta} \ell(\phi)\}, \\ \frac{\mathrm{d}p(\phi)}{\mathrm{d}\phi} &= \lambda_1 + \lambda_2 \phi^{1/\eta - 1} \ell(\phi)/\eta + o\{\phi^{1/\eta - 1} \ell(\phi)\}, \end{aligned} \tag{16}$$

for $\lambda_1, \lambda_2 \geq 0$, and $\eta < 1$, and where $\ell(\cdot)$ is a slowly-varying function at zero such that, for all $s > 0$, $\ell(s\phi)/\ell(\phi) \to 0$ as $\phi \downarrow 0$; these regularity assumptions hold for a number of popular parametric copulas (see Krupskii and Joe, 2019). We note that, when $d = 2$, η corresponds to the coefficient of tail dependence proposed by Ledford and Tawn (1996); the extension to d-dimensions is described by Eastoe and Tawn (2012).

Under the model in Eq. (16), the parameters λ_1, λ_2, and η can be estimated by defining a small threshold $\phi_* > \phi$, and considering $U_{\max} \mid U_{\max} < \phi_*$. Then, as $\phi \downarrow 0$,

$$\Pr\{U_{\max} < \phi \mid U_{\max} < \phi_*\} = p(\phi)/p(\phi_*) \sim k_1 \phi + k_2 \phi^{1/\eta}, \tag{17}$$

[1] In our application to the prediction challenge data, $F_i(\cdot)$ is known to be the standard Gumbel distribution function for all $i = 1, \ldots, d$; see Rohrbeck et al. (2023).

where $k_1 = \lambda_1/(\lambda_1\phi_* + \lambda\phi_*^{1/\eta})$ and $k_2 = (1 - k_1\phi_*)/\phi_*^{1/\eta}$. By fixing ϕ_* and assuming equality between the left and right-hand sides of Eq. (17), parameter estimates $\hat{k}_1$ and $\hat{\eta}_1$ can be computed using maximum likelihood methods. Then, our estimate of $p(\phi)$ is

$$\widehat{p(\phi)} := \{\hat{k}_1\phi + (1 - \hat{k}_1\phi_*)\phi^{1/\hat{\eta}}\}\widehat{p(\phi_*)}, \tag{18}$$

where $\widehat{p(\phi_*)}$ is the empirical estimate of $p(\phi_*)$.

The estimator in Eq. (18) can be directly used to estimate p_3 in Eq. (14) by setting $\phi = \phi_1 = 1/300$. However, it cannot be used to estimate p_4, as the formulation of the joint survival probability in Eq. (15) does not account for different levels of marginal tail decay for each component of $\mathbf{Y}$. In fact, we require an adaptation of $p(\phi)$ such that each component $U_i, i = 1, \ldots, d$, exceeds a scaled value of the form $1 - c_i\phi$ for $0 < c_i < 1/\phi$. Setting $c_i = 1$ for $i = 1, \ldots, 25$ and $c_i = 12$, otherwise, yields p_3 in Eq. (14). We construct such an estimator by considering, for $\mathbf{c} := (c_1, \ldots, c_d)'$, the probability

$$p(\phi, \mathbf{c}) := \Pr\{U_1 > 1 - c_1\phi, \ldots, U_d > 1 - c_d\phi\} = \Pr\{U_{\max}(\mathbf{c}) < \phi\}, \tag{19}$$

for weighted maxima $U_{\max}(\mathbf{c}) := \max\{(1 - U_1)/c_1, \ldots, (1 - U_d)/c_d\}$; note that when $\mathbf{c} = (1, \ldots, 1)'$, we have equivalence between Eq. (15) and Eq. (19). An estimator of the form in Eq. (19) was alluded to by Krupskii and Joe (2019); however, they did not provide theoretical results for its asymptotic behaviour. We choose to estimate $p(\phi, \mathbf{c})$ by assuming that, as $\phi \downarrow 0$, $\Pr\{U_{\max}(\mathbf{c}) < \phi\}$ has the same parametric form as $p(\phi)$ in Eq. (16). Inference for $p(\phi, \mathbf{c})$ then follows in a similar manner as for $p(\phi)$, only with samples of $U_{\max}(\mathbf{c}) \mid (U_{\max}(\mathbf{c}) > \phi_*)$ (rather than $U_{\max} \mid (U_{\max} > \phi_*)$) used for inference.

3 Results

3.1 Univariate quantile estimation

We now describe estimation of the univariate conditional quantiles required for the data prediction challenge (Rohrbeck et al., 2023). To estimate the conditional $q-$quantile of

$Y \mid (\mathbf{X} = \mathbf{x})$ for $q = 0.9999$, we fit the GPD-GAM model described in Section 2.1. The structure of $u(\mathbf{x})$, $\sigma_u(\mathbf{x})$, and $\xi(\mathbf{x})$ are optimised, for a fixed value of λ, by minimising the model's BIC for different additive combinations of linear and smooth functions of the covariates. All smooth terms are centered at zero and represented as univariate thin-plate splines, with their degrees of freedom optimised automatically using the default penalisation options available in the `R` package `evgam` (Youngman, 2022). Missing covariate values are imputed to the marginal mean of the observed values and missingness is treated as a factor variable. That is, smooth terms in a model output one of two values, depending on whether or not the input covariate is missing.

We optimise λ by using the threshold selection scheme proposed by Varty et al. (2021) and extended by Murphy et al. (2024). We first follow Heffernan and Tawn (2001) and use the quantile and GPD-GAM model estimates to transform all data onto standard exponential margins. Then, we define a grid of n equally spaced probabilities, $\{0 < \lambda_1 < ... < \lambda_n < 1\}$, with $\lambda_n = 1 - (\lambda_2 - \lambda_1)$. Let $F_E(\cdot)$ be the standard exponential distribution function and $\lambda_* \in (0, 1)$ be a pre-specified cut-off threshold. For all $i = 1, \ldots, n$, we define a vector of weights $\mathbf{w} = (w_1, \ldots, w_n)'$ with components $w_i = F_E^{-1}(\lambda_i) / \sum_{j=1}^{n} F_E^{-1}(\lambda_j)$ for $\lambda_i > \lambda_*$ and zero, otherwise. We then define the tail-weighted standardised mean absolute deviance (twsMAD) by $(1/n) \sum_{i=1}^{n} w_i |\tilde{\theta}(\lambda_i) - F_E^{-1}(\lambda_i)|$, where $\tilde{\theta}(\lambda_i)$ denotes the empirical λ_i-quantile of the standardised data. We set $\lambda_* = 0.99$ and evaluate the twsMAD for a range of candidate λ values that subceed λ_*. We then choose the optimal value which minimises this metric.

Using the BIC and twsMAD, our final model uses $\lambda = 0.972$ and, for the covariates, we include: season and wind speed as linear terms in $u(\mathbf{x})$, $\sigma_u(\mathbf{x})$, and $\xi(\mathbf{x})$; V_2, V_3, V_4, and wind direction as smooth terms in $u(\mathbf{x})$ and $\sigma_u(\mathbf{x})$. The median and the 50% confidence intervals for the test conditional q-quantiles (see Section 2.1) are estimated using the empirical quantiles over 2500 non-parametric bootstrap samples, and presented in Figure 1. The true values of the quantiles are provided by Rohrbeck et al. (2023). We observe good predictions of the conditional quantiles, with our estimates close to the true values. Our framework

attains a 38% coverage rate, which is a slight underestimation of the normal 50%.

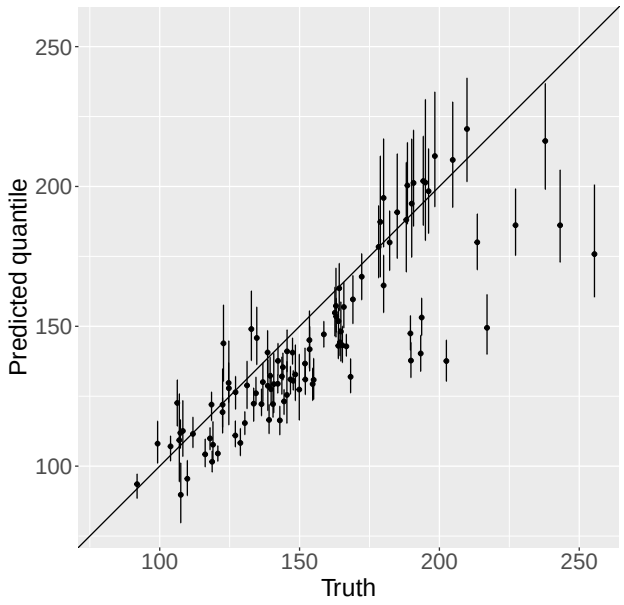

Figure 1: Predicted conditional quantiles against true values for the 100 test covariate sets for sub-challenge C1. Vertical error bars denote the 50% confidence interval, with the black points denoting the median over bootstrap samples.

To estimate the required extreme q-quantile θ of Y, we use the NBE framework detailed in Section 2.2 with $n = 21000$ independent replicates. Neural networks are trained using the Adaptive Moment Estimation (Adam) algorithm (Kingma and Ba, 2014) with a mini-batch size of 32 and, in order to improve numerical stability, input data are standardised prior to training by subtracting and dividing by the mean and standard deviation (evaluated over the entire training data set), respectively. We use $K = 350,000$ and $K/5$ parameter values for training and validation, respectively, with $n^* = 4 \times 10^7$ replicates used to estimate the theoretical q-quantiles (θ; see Section 2.2). When producing training data, we use an empirical prior distribution for the hyper-parameter set $\{(u_t, \sigma_t, \xi_t) : t = 1, \ldots, n\}$. To construct this prior, we estimate the GPD-GAM model, as described above but with $\lambda = 0.6$ (see Section 2.2 for details), across 750 bootstrap samples. Draws from the empirical prior on $\{(u_t, \sigma_t, \xi_t) : t = 1, \ldots, n\}$ then correspond to a sample from all bootstrap parameter estimates. To increase the variety of models (and hence, quantiles) used to train the estimator,

we permute the values of u_t, σ_t, and ξ_t across $\{(u_t, \sigma_t, \xi_t) : t = 1, \ldots, n\}$ when constructing the training and validation data. As small values of $Y_t < u_t$ are unlikely to impact the high q-quantile we seek to infer, we take the distribution of non-exceedances $F_t^{\leq}(\cdot)$ to be the empirical distribution of observations $\{y_t : y_t \leq u_t, t = 1, \ldots, n\}$, i.e., all observations that subceed the time-varying threshold u_t. We note that this distribution does not vary over t, but does vary with the prior draw of $\{u_t : t = 1, \ldots, n\}$.

The estimator is trained using early stopping (see, e.g., Prechelt, 2002); the validation loss is recorded at each epoch during training, which halts if the validation loss has not decreased for 10 epochs. We choose the optimal architecture for the neural networks, ϕ and $\boldsymbol{\psi}$ in Eq. (5), by minimising the risk for a test set of 1000 parameter values (not used in training); this is provided in Table 1. NBEs are trained on GPUs randomly selected

Table 1: Optimal DNN architecture used in Section 3.1. All layers used rectified linear unit (ReLU) activation functions, except the final layer, which used the identity function.

Function	**input dimension**	**output dimension**	**parameters**
$\boldsymbol{\psi}$	[1]	[48]	49
	[48]	[48]	2352
a	[48]	[48]	0
ϕ	[48]	[48]	2352
	[48]	[1]	49
total trainable parameters:			4802

from within KAUST's Ibex cluster, see `https://www.hpc.kaust.edu.sa/ibex/gpu_nodes` for details (last accessed 13/07/2023).

To illustrate the efficacy of our estimator, Figure 2 presents extreme quantile estimates for 1000 test data sets. We observe that estimates are generally aligned to the left of the diagonal, i.e., the majority of estimates in Figure 2 are overestimates. This is a consequence of the asymmetric loss function (also presented in Figure 2), which favours conservative estimates of the quantile. We estimate the q-quantile for $q = 1 - (6 \times 10^4)^{-1}$ to be 201.25. An estimated 95% confidence interval of (200.79, 201.73) was derived using a non-parametric bootstrap; due to the amortised nature of our neural estimator, non-parametric bootstrap-

based uncertainty estimation is extremely fast, as the estimator does not need to be retrained for every new sample.

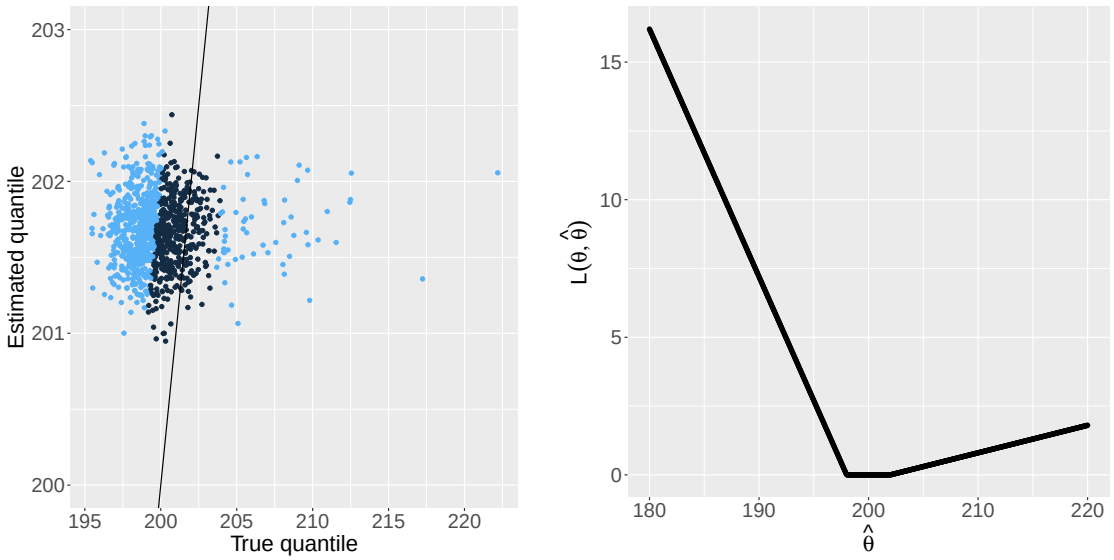

Figure 2: Left: Extreme quantile estimates for 1000 test data sets. Dark blue points denote those for which $L(\theta, \hat{\theta}) = 0$. Right: Loss function, $L(\theta_0, \hat{\theta})$ in Eq. (2), for $\theta_0 := 200$.

3.2 Extremal dependence estimation

For sub-challenge C3, we estimate p_1 and p_2 in Eq. (7) using the conditional extremes model described in Section 2.3. We perform model selection to identify the best covariates to include in the linear model for the dependence parameters, $\boldsymbol{\alpha}_{-i,t}$ and $\boldsymbol{\beta}_{-i,t}$ in Eq. (9), by minimising the AIC over the three separate model fits, i.e., $\mathbf{Y}_{-i,t} \mid (Y_{i,t} > u), i = 1, 2, 3$, with u set to the 90% standard Laplace quantile. The best fitting model did not include any transformation (e.g, log or square) of the covariates, and uses a homogeneous $\boldsymbol{\beta}_{-i,t}$ function, i.e., with $\boldsymbol{\beta}_{-i,t} = \boldsymbol{\beta}_{-i}$ for all $t = 1, \ldots, n$, and for $i = 1, 2, 3$.

We use a visual diagnostic to optimise the exceedance threshold u. We fit the three conditional models for a grid of u values, use Algorithm 2 (with $N' = 5N$) to simulate $N := 10^7$ realisations from the fitted model for $\mathbf{Y} = (Y_1, Y_2, Y_3)$, i.e., with time t marginalised out, and then compute realisations of $R := Y_1 + Y_2 + Y_3$. The upper tails of the aggregate variable R are heavily influenced by the extremal dependence structure in the underlying

Table 2: Sub-challenge C3: Parameter estimates $\hat{\boldsymbol{\alpha}}^{(0)}_{-i}, \hat{\boldsymbol{\alpha}}^{(1)}_{-i}$, and $\hat{\boldsymbol{\beta}}_{-i}$, for the conditional extremal dependence model, $\mathbf{Y}_{-i,t} \mid Y_{i,t} > u$, $i = 1, 2, 3$. Here, $\hat{\boldsymbol{\alpha}}^{(1,j)}_{-i}$ denotes the j^{th} column of $\hat{\boldsymbol{\alpha}}^{(1)}_{-i}$ with $j = 1$ and $j = 2$ corresponding to the linear coefficient vector for *Season* and *Atmosphere*, respectively. Values in the j^{th} row give the component of the parameter vector for $\mathbf{Y}_{-i,t}$ that corresponds to $Y_{j,t}$, $j = 1, 2, 3$ (and so are missing for $j = i$).

	$i = 1$				$i = 2$				$i = 3$			
	$\hat{\boldsymbol{\alpha}}^{(0)}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(1,1)}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(1,2)}_{-i}$	$\hat{\boldsymbol{\beta}}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(0)}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(1,1)}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(1,2)}_{-i}$	$\hat{\boldsymbol{\beta}}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(0)}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(1,1)}_{-i}$	$\hat{\boldsymbol{\alpha}}^{(1,2)}_{-i}$	$\hat{\boldsymbol{\beta}}_{-i}$
$Y_{1,t}$	-	-	-	-	-0.08	0.22	0.08	0.81	0.15	-0.33	0.01	0.97
$Y_{2,t}$	0.56	-0.06	0.07	0.70	-	-	-	-	0.01	-0.07	0.01	0.45
$Y_{3,t}$	0.28	-0.02	0.07	0.82	0.24	-0.14	0.14	0.09	-	-	-	-

(Y_1, Y_2, Y_3) (Richards and Tawn, 2022). Hence, we can exploit the empirical distribution of the aggregate as a diagnostic measure for the quality of the dependence model fit. We choose the optimal u as that which provides the best estimates for extreme quantiles of R; this is achieved through visual inspection of a Q-Q plot (see, also, Richards et al., 2022, 2023c). For brevity, we test only two values of u: the 90% and 95% standard Laplace quantiles. We found the optimal u to be the latter, and we illustrate the corresponding Q-Q plot in Figure 3. Table 2 provides the corresponding parameters estimates $\hat{\boldsymbol{\alpha}}^{(0)}_{-i}, \hat{\boldsymbol{\alpha}}^{(1)}_{-i}$, and $\hat{\boldsymbol{\beta}}_{-i}$, for $i = 1, 2, 3$.

Example realisations from the fitted model are illustrated in Figure 3. Good model fit is illustrated, as the realisations mimic the extremal characteristics of the observed data. Probabilities p_1 and p_2 are estimated empirically from the aforementioned $N = 10^7$realisations of (Y_1, Y_2, Y_3), and found to be $\hat{p}_1 = 3.13 \times 10^{-5}$ and $\hat{p}_2 = 2.29 \times 10^{-5}$, respectively.

For sub-challenge C4, we estimate p_3 and p_4 in Eq. (14) using the non-parametric tail probability estimators defined in Eq. (18) and Eq. (19), respectively. However, we first seek to decompose the high-dimensional random vector $\mathbf{Y}$ into strongly tail-dependent sub-vectors. Our reasoning is threefold: i) the estimator in Eq. (19) requires strong assumptions about the joint upper tail decay of the random vector $\mathbf{Y}$ which are unlikely to hold for high dimension d; ii) reliable estimation of the intermediate exceedance probability, $\widehat{p(\phi_*)}$ in Eq. (18), becomes difficult as d grows large; iii) Krupskii and Joe (2019) observed through

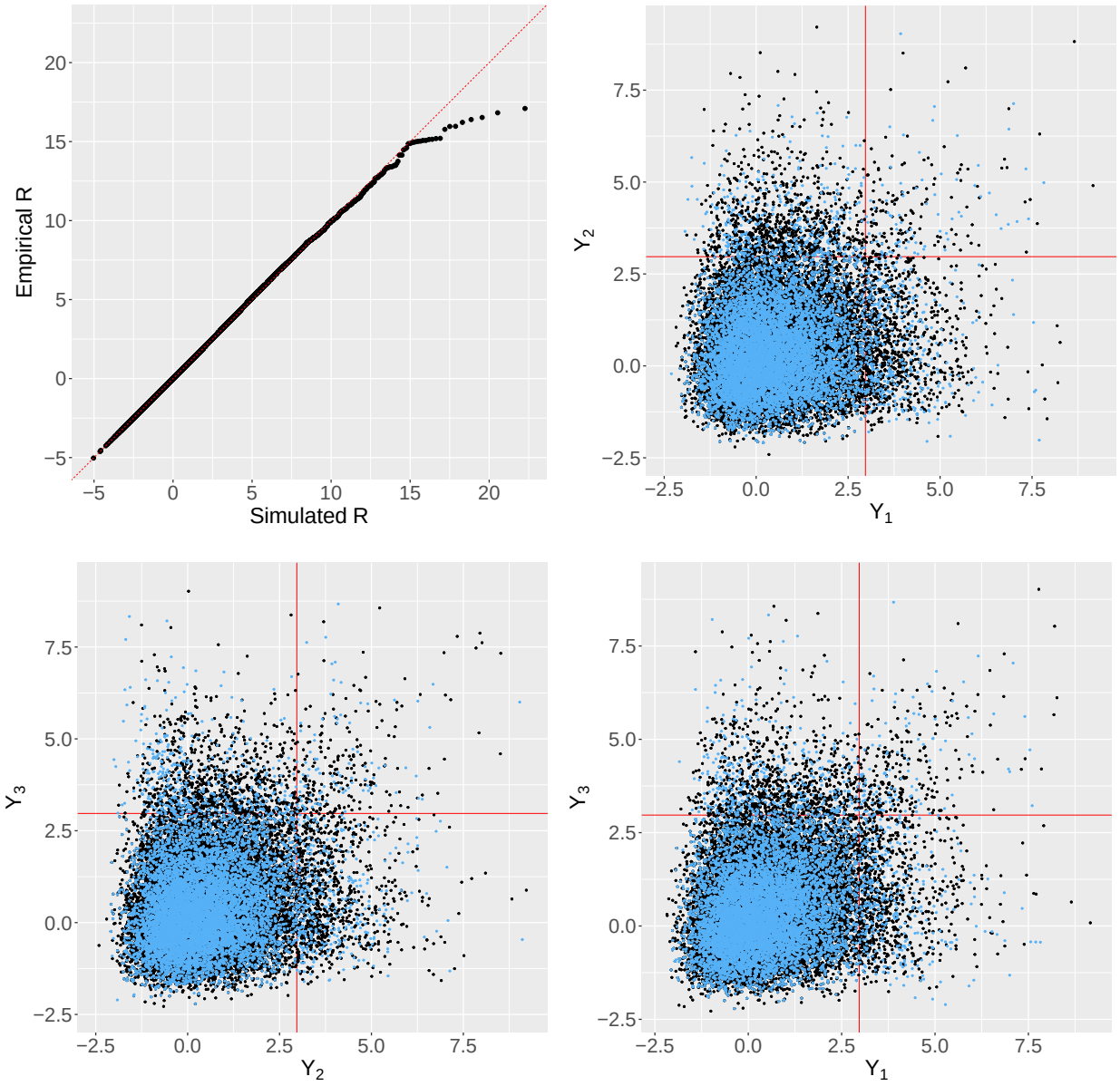

Figure 3: Top-left panel: Q-Q plot for the simulated and empirical sum $R = Y_1 + Y_2 + Y_3$. Other panels provide scatter plots of (black) observations and (blue) realisations from the fitted conditional extremal dependence models. Red vertical and horizontal lines correspond to the exceedance threshold u, i.e., the 95% marginal quantile.

simulations that the estimator in Eq. (18) provides more accurate estimates when the true value of η is close to one, i.e., when $\mathbf{Y}$ exhibits strong upper tail dependence.

We identify sub-vectors by estimating the EDM between all pairs of components of $\mathbf{Y}$, with $\alpha = 2$, $\|\cdot\|$ set to the L_2 norm, and with u in Eq. (13) taken to be the empirical

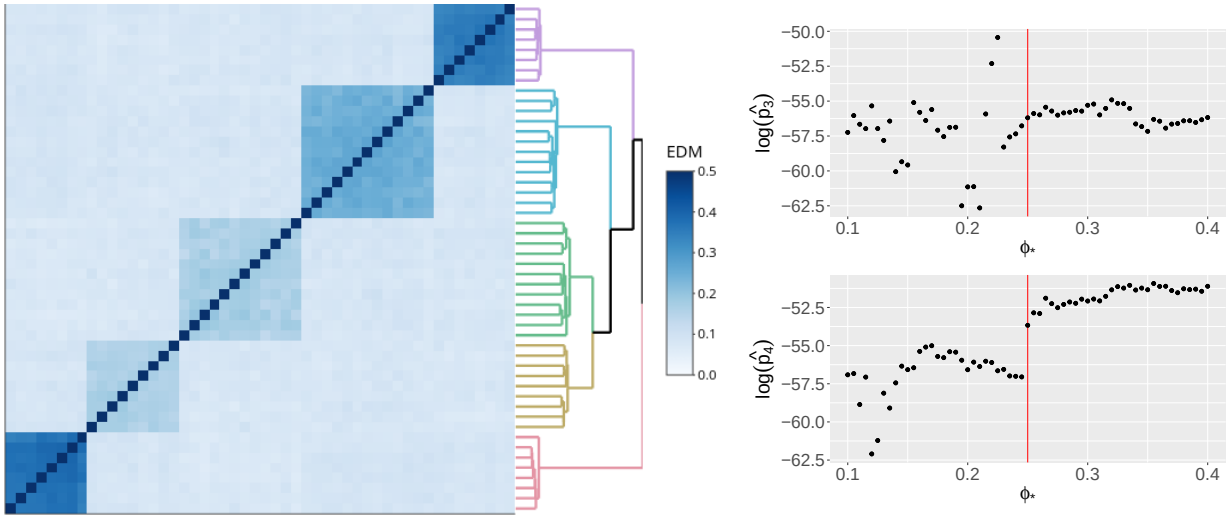

Figure 4: Left: Heatmap of pairwise EDM estimates grouped using hierarchical clustering, alongside the corresponding dendrogram. Right: Threshold stability plots of estimated log-exceedance probabilities, $\log(\hat{p_3})$ and $\log(\hat{p_4})$, against the intermediate quantile ϕ_* used in their estimation. Red horizontal lines denote the optimal choice of ϕ_*.

0.99-quantile. Estimates are presented in a heatmap in Figure 4, with values grouped using hierarchical clustering. An appropriate grouping can be found by visual inspection of Figure 4; the clustered heatmap produces a block-diagonal matrix with five blocks (with size ranging from 8–13 components), and with elements of the off-diagonal blocks close to zero. For less well-behaved data, the number of clusters (and sub-vectors of $\mathbf{Y}$) can instead be determined by placing a cut-off at an appropriate point on the clustering dendrogram.

We decompose $\mathbf{Y}$ into the five sub-vectors suggested by Figure 4, which we denote by $\mathcal{Y}_i$ for $i = 1, \ldots, 5$. Estimation of p_3 then follows under the approximation

$$p_3 = \Pr(Y_1 > s_1, \ldots, Y_{50} > s_1) \approx \prod_{i=1}^{5} \Pr\left(\bigcap_{Y \in \mathcal{Y}_i} Y > s_1 \right), \tag{20}$$

and similarly for p_4 (with the exceedance value s_1 or s_2 chosen appropriately). That is, we assume complete independence between the sub-vectors of $\mathbf{Y}$ to approximate p_3 and p_4. Exceedance probabilities for each sub-vector are estimated using the methodology described in Section 2.5. We fix the intermediate quantile level ϕ_* in Eq. (18) across all sub-vectors and for estimation of both p_3 and p_4. An optimal value $\hat{\phi}_*$ is chosen via a threshold stability plot (see, e.g., Coles, 2001). In Figure 4, we present estimates of $\log(p_3)$ and $\log(p_4)$ for a

sequence of ϕ_* values in $[0.1, 0.4]$. We choose $\hat{\phi}_*$ as the smallest ϕ_* such that there is visual stability in Figure 4 for estimates of both $\log(p_3)$ and $\log(p_4)$ when $\phi_* > \hat{\phi}_*$. In this case, we adopt $\hat{\phi}_* = 0.25$ as the optimal value. We use a non-parametric bootstrap with 200 samples to assess uncertainty in estimates of p_3 and p_4. The resulting bootstrap median[2] estimates of $\log(p_3)$ and $\log(p_4)$ (and their estimated 95% confidence intervals) are -59.38 $(-68.46, -49.82)$ and -55.24 $(-60.12, -49.79)$, respectively.

4 Conclusion

We proposed four frameworks for the estimation of exceedance probabilities and levels associated with extreme events. To estimate extreme conditional quantiles, we adopted an additive model that represents parameters in a peaks-over-threshold model as splines. To estimate univariate quantiles in an unconditional setting, we constructed a neural Bayes estimator that estimates a quantile subject to a conservative loss function. Our new approach to quantile estimation is amortised, likelihood-free, and requires few parametric assumptions about the underlying distribution. For multivariate data, we considered two frameworks: i) in the presence of covariates, we adopted a non-stationary conditional extremal dependence model to capture linear trends in extremal dependence parameters, and ii) in the absence of covariates, we proposed a non-parametric estimator for multivariate tail probabilities that can be applied to high-dimensional ($d = 50$) data via a sparse decomposition using pairwise extremal dependence measures. We validated these modelling approaches by using them to win the EVA (2023) Conference Data Challenge (Rohrbeck et al., 2023).

We illustrated the efficacy of additive GPD regression models for estimating extreme conditional quantiles. Machine learning-based extreme quantile regression models using, for example, random forests (Gnecco et al., 2024), gradient boosting (Velthoen et al., 2023), and neural networks (Pasche and Engelke, 2022; Richards et al., 2023a; Richards and Huser, 2024) may offer a potentially more flexible alternative to additive models.

[2]We submitted the median over bootstrap estimates for the data competition.

We designed a neural Bayes estimator (NBE) to perform simulation-based inference for extreme quantiles. To train this estimator, we constructed an empirical prior measure $\pi(\cdot)$ for the extreme quantile, by simulating from a family of pre-defined peaks-over-threshold models. The resulting NBE was optimal with respect to the prior measure $\pi(\cdot)$. Whilst our estimator performed well in the application, the underlying truth was a peaks-over-threshold model (Rohrbeck et al., 2023), and, hence, the truth distribution was contained within the class of distributions covered by $\pi(\cdot)$. We note that our estimator may not perform as well when the converse holds, and the true distribution of Y is not approximated well by any family in $\pi(\cdot)$. Further study of the optimality properties of our proposed NBE is required and, in particular, its efficacy under model misspecification.

We proposed an extension of the conditional extremal dependence regression model (Winter et al., 2016) and highlighted its efficacy for estimation of extreme exceedance probabilities. Our conditional model performed well in practice, but alternative multivariate extremal dependence regression models, which rely on more classical assumptions about the extremal behaviour of $\mathbf{Y}$, e.g., regular or hidden regular variation, could have been tested (see, e.g., Cooley et al., 2012; de Carvalho et al., 2022; Murphy-Barltrop and Wadsworth, 2022). In our application, we tested only two values for the exceedance threshold u in (8), and quantified goodness-of-fit through visual inspection of Q-Q plots for the aggregate R (see Figure 3). A more rigorous testing procedure could be employed, whereby a grid of u values are tested and the optimal u is taken to minimise some numerical measure of fit; see, e.g., Murphy et al. (2024), who proposed an automated threshold selection procedure for peaks-over-threshold models.

Finally, we proposed an extension of the non-parametric tail probability estimator of Krupskii and Joe (2019) to account for different levels of marginal decay for components of a random vector. Whilst we were able to showcase the efficacy of our estimator by using it to win sub-challenge C4 of the EVA (2023) Conference Data Challenge, we have not provided any theoretical guarantees for the estimator; we leave this as a future endeavour.

Declarations

Funding

All authors were supported by funding from the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2020-4394.

Acknowledgments

The authors are grateful to Raphaël Huser and Matthew Sainsbury-Dale for helpful discussions and to Jonathan Tawn, Christian Rohrbeck, and Emma Simpson of Lancaster University, University of Bath, and University College London, respectively, for organisation of the data challenge that motivated this work. Support from the KAUST Supercomputing Laboratory is gratefully acknowledged.

References

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98.

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(1):207–222.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, volume 208. Springer, London.

Cooley, D., Davis, R. A., and Naveau, P. (2012). Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. *The Annals of Applied Statistics*, 6(4):1406–1429.

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 52(3):393–425.

de Carvalho, M., Kumukova, A., and Dos Reis, G. (2022). Regression-type analysis for multivariate extreme values. *Extremes*, 25(4):595–622.

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 Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8:241–270.

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.

Gerber, F. and Nychka, D. (2021). Fast covariance parameter estimation of spatial Gaussian process models using neural networks. *Stat*, 10(1):e382.

Gnecco, N., Terefe, E. M., and Engelke, S. (2024). Extremal random forests. *Journal of the American Statistical Association*, pages 1–14.

Heffernan, J. E. (2000). A directory of coefficients of tail dependence. *Extremes*, 3:279–290.

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. (2001). Extreme value analysis of a large designed experiment: a case study in bulk carrier safety. *Extremes*, 4:359–378.

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: Methodology*, 66(3):497–546.

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

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.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krupskii, P. and Joe, H. (2019). Nonparametric estimation of multivariate tail probabilities and tail dependence coefficients. *Journal of Multivariate Analysis*, 172:147–161.

Larsson, M. and Resnick, S. I. (2012). Extremal dependence measure and extremogram: the regularly varying case. *Extremes*, 15(2):231–256.

Lederer, J. and Oesting, M. (2023). Extremes in high dimensions: methods and scalable algorithms. *arXiv preprint arXiv:2303.04258*.

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

Lenzi, A., Bessac, J., Rudi, J., and Stein, M. L. (2023). Neural networks for parameter estimation in intractable models. *Computational Statistics & Data Analysis*, 185:107762.

Lenzi, A. and Rue, H. (2023). Towards black-box parameter estimation. *arXiv preprint arXiv:2303.15041*.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, 30(5):1225–1310.

Murphy, C., Tawn, J. A., and Varty, Z. (2024). Automated threshold selection and associated inference uncertainty for univariate extremes. *arXiv preprint arXiv:2310.17999*.

Murphy-Barltrop, C. and Wadsworth, J. (2022). Modelling non-stationarity in asymptotically independent extremes. *arXiv preprint arXiv:2203.05860*.

Pasche, O. C. and Engelke, S. (2022). Neural networks for extreme quantile regression with an application to forecasting of flood risk. *arXiv preprint arXiv:2208.07590*.

Prechelt, L. (2002). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, New York.

Rai, S., Hoffman, A., Lahiri, S., Nychka, D. W., Sain, S. R., and Bandyopadhyay, S. (2023). Fast parameter estimation of generalized extreme value distribution using neural networks. *arXiv preprint arXiv:2305.04341*.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5:303–336.

Resnick, S. (2004). The extremal dependence measure and asymptotic independence. *Stochastic Models*, 20(2):205–227.

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media, New York.

Richards, J. and Huser, R. (2024). Regression modelling of spatiotemporal extreme U.S. wildfires via partially-interpretable neural networks. *arXiv preprint arXiv:2208.07581*.

Richards, J., Huser, R., Bevacqua, E., and Zscheischler, J. (2023a). Insights into the drivers and spatio-temporal trends of extreme Mediterranean wildfires with statistical deep-learning. *Artificial Intelligence for the Earth Systems*, 2(4):e220095.

Richards, J., Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2023b). Neural Bayes estimators for censored inference with peaks-over-threshold models. *arXiv preprint arXiv:2306.15642*.

Richards, J. and Tawn, J. A. (2022). On the tail behaviour of aggregated random variables. *Journal of Multivariate Analysis*, 192:105065.

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.

Richards, J., Tawn, J. A., and Brown, S. (2023c). Joint estimation of extreme spatially aggregated precipitation at different scales through mixture modelling. *Spatial Statistics*, 53:100725.

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):246–282.

Rohrbeck, C., Simpson, E. S., and Tawn, J. A. (2023). Editorial: EVA 2023 data challenge.

Sainsbury-Dale, M., Richards, J., Zammit-Mangion, A., and Huser, R. (2023). Neural bayes estimators for irregular spatial data using graph neural networks. *arXiv preprint arXiv:2310.02600*.

Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2024). Likelihood-free parameter estimation with neural bayes estimators. *The American Statistician*, 78(1):1–14.

Shooter, R., Tawn, J., Ross, E., and Jonathan, P. (2021). Basin-wide spatial conditional extremes for severe ocean storms. *Extremes*, 24:241–265.

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*.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2023). Gradient boosting for extreme quantile regression. *Extremes*, pages 1–29.

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

Winter, H. C., Tawn, J. A., and Brown, S. J. (2016). Modelling the effect of the El Niño-Southern Oscillation on extreme spatial temperature events over Australia. *The Annals of Applied Statistics*, 10(4):2075 – 2101.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for US wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.

Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software*, 103(3):1–26.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach. Curran Associates, Inc.

Zammit-Mangion, A. and Wikle, C. K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spatial Statistics*, 37:100408.