Editorial: EVA (2023) Conference Data Challenge

Christian Rohrbeck[1], Emma S. Simpson[2], Jonathan A. Tawn[3*]

[1]Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, U.K.
[2]Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, U.K.
[3]School of Mathematical Sciences, Lancaster University, Bailrigg, LA1 4YF, U.K.

*Corresponding author(s). E-mail(s): j.tawn@lancaster.ac.uk;
Contributing authors: cr777@bath.ac.uk; emma.simpson@ucl.ac.uk;

1 Introduction

It was a pleasure to be entrusted with setting the EVA 2023 Data Challenge and we were really pleased to have received entries from seven very committed teams, which in alphabetical order by team name are genEVA (Geneva), Lancopula Utopiversity (Lancaster and Maynooth), SHSmultiscale (Sejong and Seoul), Uniofbathtopia (Bath), Wee_Extremes (Glasgow), Yahabe (Montréal and Kanpur) and Yalla (KAUST). Here, the names in parentheses are the institutions where the team members are from, with teams not being formal representatives of these institutions.

In designing the challenge, we decided to create problems that capture the variety of contexts we have experienced in the analysis of environmental extremes data. Specifically, we wanted to replicate the role of an applied statistician with the need to estimate quantiles and probabilities for extreme events in univariate and multivariate settings, driven by end-user considerations. We set four challenges, labelled C1-C4, with the first (last) two focusing on univariate (multivariate) analyses. These four components were assessed and ranked separately over the teams, before being combined into a single score and ranking.

The univariate extremes problems involve inference for extreme quantiles, with teams being asked to obtain point estimates and reliable, useful confidence intervals. However, with environmental data we are faced with additional complications such as

covariates; data missing completely at random; and the need to convert the inference into design levels which account for different losses from over- and under-design.

In the multivariate extremes problems, we wanted to assess the teams' performance in a way that focused entirely on the dependence modelling component. Consequently, the multivariate problems relate to data where the univariate marginal distributions are all known. Here, the complexity comes from estimating probabilities of extreme events in different dimensions and with respect to different marginal quantiles.

This editorial for the data challenge is structured as follows: Section 2 presents details of the data, covariates and the information provided to the teams about the true generating processes of the data; Section 3 presents the challenges we set, with some clarifications and minor modifications that were introduced during the course of the challenge, as well as details of the assessment methods. In Section 4, we present our underlying stochastic models and computational details of how we derived the true values for each challenge. Section 5 summarizes the performances of the seven teams that entered the competition. Finally, in Section 6, we provide an overview of the methods described in the papers published in this special issue and discuss their suitability with respect to the four challenges.

2 Data, Covariates and Known Properties

Our data come from a rather unique and special country, called Utopia, where everything is a bit idealised. We have data from its capital city, Amaurot, and from two of its islands, Coputopia and Utopula. We are interested in extreme values of the environmental variable denoted by Y, with the notation $Y_{i,t}$ identifying the variable Y at site i on day t. On each individual island, there may be dependence between the Y variables at different sites on that island, but they are known to be independent of the variables Y on the other island and at Amaurot.

For Amaurot and Coputopia there are 70 years of daily data, but only 50 years for Utopula. In Utopia, spatial information is irrelevant and no knowledge of any environmental process on Earth is applicable. However, like Earth, Utopia suffers from data recording problems, with 11.7% of the provided observations for Amaurot having at least one missing value, which it is reasonable to assume are missing completely at random.

For each day, we have a vector of covariates $\boldsymbol{X}_t = (V_{1,t}, \ldots, V_{8,t})$, with (V_5, V_6, V_7, V_8) corresponding to (season, WS = windspeed, WD = wind direction, atmosphere), but with the other covariates (V_1, V_2, V_3, V_4) unnamed. Given the covariates, the $\{Y_{i,t}\}$ are independent over t for each given i. Utopia has experienced a very stable climate over the observation period, and experts predict that this won't change in the next decades, so it is reasonable to assume that observed covariate patterns in the data are representative over all time periods - WS and WD data exhibited non-stationarity, but it was communicated to the teams that this property was created unintentionally (the data should have been additionally randomised over these features) and so should be ignored. Only season and atmosphere have a changing structure (over 1 and 70 years, respectively), in a cyclically repeating pattern. In Utopia, each year consists of 12 months with 25 days each, and it is equally split into seasons 1 and 2.

On each of the islands, the marginal distributions are known to be identical across all sites and over time, with standard Gumbel distributions. On Coputopia there are three sites, and teams were provided with information on season and atmospheric conditions. In contrast, on Utopula there are 50 sites, split equally over two regions U1 and U2, for which the joint distribution of $Y_{i,t}$ across sites is identical over time.

3 The Challenges and their Assessment

In this section, we summarise the four challenges that the teams were asked to complete.

C1: For Amaurot, build a model for the distribution of $Y \mid \boldsymbol{X}$ and estimate the 0.9999-quantile of the conditional distribution and associated 50% confidence intervals for 100 different provided covariate combinations. Specifically, the quantiles $\{q(\boldsymbol{x}_i) : i = 1, \ldots, 100\}$ satisfy

$$\Pr(Y < q(\boldsymbol{x}_i) \mid \boldsymbol{X} = \boldsymbol{x}_i) = 0.9999,$$

for covariate combinations $\{\boldsymbol{x}_i : i = 1, \ \ldots, 100\}$.

Here, the assessment of a team's performance was based on the accuracy of the actual coverage of their confidence intervals over the 100 produced intervals, i.e., how close they are to having 50% coverage of the true values. To avoid any teams gaming the problem, e.g., by producing 50 very narrow intervals and 50 very wide intervals, we also looked at the proximity of their point estimates to the truth and asked for their code to verify they were implementing a genuine statistical algorithm for the confidence interval construction.

C2: Again for Amaurot, estimate the marginal quantile q such that $\Pr(Y > q) = (6 \times 10^4)^{-1}$. This quantile corresponds to a once in 200 years level, if the process was independent and identically distributed.

As this quantile estimate is to be used for design purposes, the challenge is to account for the potential losses that could be incurred from over- or under-estimating q. Over-estimating would mean more has to be spent to protect against Y than necessary. Under-estimating q would lead to more regular environmental damage to Amaurot than is expected, thus resulting in large insurance claims. A small error in the estimate $\hat{q}$ of q is acceptable but an under-estimation is considered worse than an equal level of over-estimation. So, quantile inference should minimise the loss function

$$L(q, \hat{q}) = \begin{cases} 0.9\,(0.99q - \hat{q}), & \text{if } 0.99q > \hat{q}, \\ 0, & \text{if } |q - \hat{q}| \leq 0.01q, \\ 0.1\,(\hat{q} - 1.01q), & \text{if } 1.01q < \hat{q}. \end{cases} \tag{1}$$

Assessment in this case is based simply on the team with the smallest loss function value $L(q, \hat{q})$ for their selected $\hat{q}$. In the case of multiple teams having $L(q, \hat{q}) = 0$, the value of $|q - \hat{q}|$ provides a secondary ranking rule, where smaller values are preferable.

C3: For the three towns on the island of Coputopia, estimate the probabilities p_1 and p_2 below, corresponding to a combination of extreme and non-extreme events simultaneously:

$$p_1 = \Pr(Y_1 > 6, Y_2 > 6, Y_3 > 6); \qquad p_2 = \Pr(Y_1 > 7, Y_2 > 7, Y_3 < m),$$

where $m = -\log(\log 2)$ is the median of Y_3, having a standard Gumbel distribution.

The assessment for both C3 and C4 involved a probability-based scoring rule taken from Smith (1999). Specifically, for the two required estimates $(\hat{p}_1, \hat{p}_2)$ we used the metric

$$P_{12} = \sum_{i=1}^{2} \left| p_i \log\left(p_i/\hat{p}_i\right) + (1-p_i) \log\left[(1-p_i)/(1-\hat{p}_i)\right] \right|, \tag{2}$$

with smaller values of P_{12} being better.

C4: For Utopula, sites $i_1 = 1, \ldots, 25$ are in U1 and sites $i_2 = 1, \ldots, 25$ are in U2. The current design standards give greater protection for sites in U1 than in U2. Specifically, let s_1 (s_2) denote the marginal level exceeded once in a year (in a month) on average. Then, the associated marginal exceedance probabilities are $\phi_2 = 12 \times \phi_1$ with $\phi_1 = 1/300$, and so $s_1 = -\log\{-\log(1-\phi_1)\}$ and $s_2 = -\log\{-\log(1-\phi_2)\}$. Estimate the joint probability

$$p_1 = \Pr(Y_{i_j,t} > s_j : i_j = 1, \ldots, 25; j = 1, 2).$$

Now suppose the design standard is made uniform across the island, with U2 standards raised to those of U1. Estimate the updated joint probability

$$p_2 = \Pr(Y_{i_j,t} > s_1 : i_j = 1, \ldots, 25; j = 1, 2).$$

4 Underlying Truth

4.1 Truth for C1 and C2

The covariates $\boldsymbol{X}_t$ were generated according to the following joint model:

- The variables V_1 and V_2 were simulated as $V_1 = W + \epsilon_1$ and $V_2 = W + \epsilon_2$, where the distribution of W is a mixture of the two normal random variables $W^{(1)} \sim \text{Normal}(30, 9)$ and $W^{(2)} \sim \text{Normal}(36, 6.25)$, with about 40% of observations being drawn from $W^{(1)}$. The variables W, ϵ_1 and ϵ_2 are independent, and $\epsilon_1 \sim \text{Normal}(0, 4)$ and $\epsilon_2 \sim \text{Gamma}(1.2, 0.3)$.
- The distribution of the variable V_3 varied across the two seasons. For season 1, a skewed-normal distribution, restricted to the positive real line, with location $\xi_{V_3}^{(1)} = 4$, scale $\omega_{V_3}^{(1)} = 4$ and shape $\alpha_{V_3}^{(1)} = 5$, was employed. Values for season 2 were generated using a generalized extreme value (GEV) distribution, $\text{GEV}(\mu, \sigma, \xi)$, with cumulative distribution function (CDF)

$$G(z \mid \mu, \sigma, \xi) = \exp\left[-\left\{ 1 + \xi \left(\frac{z-\mu}{\sigma} \right) \right\}_{+}^{-1/\xi} \right], \qquad z \in \mathbb{R}, \tag{3}$$

where $\{x\}_+ = \max(x, 0)$. Specifically, we sampled from GEV$(4, 4, 0.2)$ restricted to the positive real line, that is, the CDF of V_3 for season 2 is given as

$$F_{V_3}(v) = \frac{G(v \mid \mu = 4, \sigma = 4, \xi = 0.2) - G(0 \mid \mu = 4, \sigma = 4, \xi = 0.2)}{1 - G(0 \mid \mu = 4, \sigma = 4, \xi = 0.2)}, \qquad v > 0.$$

- Values for V_4 were drawn from a GEV(0,1,0.1), with any sampled negative values set to zero. As such, using the notation in expression (3), the CDF of V_4 is given as

$$F_{V_4}(v) = \begin{cases} 0 & \text{for } v < 0 \\ G(v \mid \mu = 0, \sigma = 1, \xi = 0.1) & \text{for } v \geq 0 \end{cases}$$

- The distribution of the wind direction was based on a two-component mixture of von-Mises distributions on the unit sphere $\mathbb{S}^1 = \{\boldsymbol{w} \in \mathbb{R}^2 : w_1^2 + w_2^2 = 1\}$, with the mixture probability set to 40%. The density of the von-Mises distribution is

$$h(\boldsymbol{w}; \boldsymbol{\mu}, \kappa) = c_0(\kappa) \exp\left(\kappa \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}\right)$$

  for $\boldsymbol{w} \in \mathbb{S}^1$, where $\boldsymbol{\mu} \in \mathbb{S}^1$ and $\kappa > 0$ are termed the mean direction and concentration parameter, respectively, and $c_0(\kappa)$ is a normalising constant. The parameters were set to $\boldsymbol{\mu} = (0.3, 0.5)$ and $\kappa = 2$ for the first mixture component, and $\boldsymbol{\mu} = (-0.4, -0.4)$ and $\kappa = 4$ for the second mixture component. Given a sampled value $\boldsymbol{w} = (w_1, w_2)$, the value for WD was defined as $\arctan(w_2/w_1)$.
- Values for wind speed were sampled from a normal distribution restricted to the positive real line with parameters $\mu_{\mathrm{WS}} = 3 + 2\sin(\mathrm{WD})$ and $\sigma_{\mathrm{WS}} = 1$.
- Finally, we used monthly values of the North Atlantic Oscillation (NAO) index for 1950-2019 to generate the values of the atmosphere variable, applying linear detrending to the mean such that the resulting covariate values are zero-centred (and therefore reasonable to repeat over a 70-year cycle).

The response variable Y_t depends both on the covariates $\boldsymbol{X}_t$ and a latent variable Z_t, where we drop the index t in the following. Specifically, the distribution of Z conditional on $\boldsymbol{X}$ is described by the generalized Pareto distribution (GPD),

$$Z \mid (\boldsymbol{X} = \boldsymbol{x}) \sim \mathrm{GPD}\left(\sigma(\boldsymbol{x})/10, -\pi^{-2}\right), \tag{4}$$

where $\sigma(\boldsymbol{x}) = \exp(-1) + 18.7\sqrt{V_2} + 9\left[1 + \log(V_3)\right]^2 + 5.71\,\mathrm{WS}^{1.5}$. The CDF of the GPD$(\sigma, \xi)$ is formally given by

$$H(z \mid \sigma, \xi) = 1 - \left(1 + \xi \frac{z}{\sigma}\right)_+^{-1/\xi},$$

where $\{z\}_+ = \max(z, 0)$. Since $\xi < 0$ in (4), $Z \mid (\boldsymbol{X} = \boldsymbol{x})$ has a bounded upper tail.

To generate observations for the random variable $Y \mid (\boldsymbol{X} = \boldsymbol{x})$ considered in the data challenge, we used rejection sampling which gives that $Y \mid (\boldsymbol{X} = \boldsymbol{x})$ and $Z \mid (\boldsymbol{X} = \boldsymbol{x})$ have the same tail distribution above a high threshold. We set the thresholds to

$u_1 = 112 - 6|\text{WD}|$ and $u_2 = 110 - 5|\text{WD}|^{0.9}$ for seasons 1 and 2 respectively. If the sampled value z for $Z \mid (\boldsymbol{X} = \boldsymbol{x})$ exceeds the associated threshold, the observation is kept as an observation for $Y \mid (\boldsymbol{X} = \boldsymbol{x})$. Otherwise, we keep z as an observation for $Y \mid (\boldsymbol{X} = \boldsymbol{x})$ with probability sampled from a Beta(z, u_1) distribution for season 1, and from a Beta$(\exp\{2 + (z - u_2)/30\}, 1)$ for season 2. Consequently, the distribution of $Y \mid (\boldsymbol{X} = \boldsymbol{x}, Y > 112)$ corresponds to a GPD, with the scale parameter dependent on three explanatory variables, while $Y \mid (\boldsymbol{X} = \boldsymbol{x})$ depends on a larger set of explanatory variables and is not GPD.

For the 100 values of $\boldsymbol{X}$ used in the validation set of C1, we sampled 50 data points from the joint distribution of $\boldsymbol{X}$, and 50 data points from the joint tail. More specifically, we sampled ten data points each from $\boldsymbol{X} \mid [\text{V}_j > q_{0.99}(\text{V}_j)]$ $(j = 1, \ldots, 4)$ and ten data points from $\boldsymbol{X} \mid [\text{WS} > q_{0.99}(\text{WS})]$, where $q_p(V)$ is the pth marginal sample quantile of covariate V.

Monte Carlo methods were used to find the true quantiles considered in the challenge questions. For C1, we sampled 30×10^6 realisations of Z for each validation point $\boldsymbol{x}_i$ $(i = 1, \ldots, 100)$, which gave between 1.5×10^6 and 10×10^6 values for $Y \mid (\boldsymbol{X} = \boldsymbol{x}_i)$. This sample was used to obtain an estimate $\hat{p}_i$ for the exceedance probability $\Pr(Y > 112 \mid \boldsymbol{X} = \boldsymbol{x}_i)$; we considered this threshold because $(Y - 112) \mid (Y > 112 \mid \boldsymbol{X} = \boldsymbol{x}_i)$ is GPD. Since C1 asks for the 99.99% quantile of $Y \mid (\boldsymbol{X} = \boldsymbol{x}_i)$, we first consider whether $\hat{p}_i > 0.0001$ or $\hat{p}_i \leq 0.0001$. For $\hat{p}_i > 0.0001$, the 99.99% quantile of $(Y - 112) \mid (\boldsymbol{X} = \boldsymbol{x}_i)$ corresponds to the $(\hat{p}_i - 0.0001)/\hat{p}_i \times 100\%$ quantile of a GPD $\left(\sigma(\boldsymbol{x}_i)/10 - 112/\pi^2, -1/\pi^2\right)$, with $\sigma(\boldsymbol{x}_i)$ as defined in (4), and this was our estimate for the quantiles in C1. For the small number of the validation points with $\hat{p}_i \leq 0.0001$, we used the empirical quantile estimate for $Y \mid (\boldsymbol{X} = \boldsymbol{x}_i)$, which we obtained from our sample from the distribution of $Y \mid (\boldsymbol{X} = \boldsymbol{x}_i)$.

To find the true quantile for C2, we sampled 50×10^6 data points for $\boldsymbol{X}$ and generated a value for $Y \mid (\boldsymbol{X} = \boldsymbol{x})$ for each sampled combination of $\boldsymbol{X}$. The estimate used for evaluating the predictions was set to the empirical quantile of these 50×10^6 sampled values for Y. Our simulation gave the true quantile of 196.6 to 4 significant figures.

4.2 Truth for C3

In this challenge, the joint distribution of (Y_1, Y_2, Y_3) is obtained via the max-mixture construction defined in (5) below (see also Simpson et al. (2020), for example). To introduce some complexity into the joint tail behaviour of the variables, the mixture components are chosen to exhibit both asymptotic dependence and asymptotic independence, and some of the dependence parameters and mixing probabilities are functions of the atmosphere covariate, V_8.

First, for each time t, we generate observations from three bivariate extreme value (BEV) copulas with logistic models (Gumbel, 1960) on Fréchet margins. The strength of dependence in these models is controlled by a constant parameter $\alpha \in (0, 1]$, with values closer to zero resulting in stronger dependence. Specifically, we take

$$\boldsymbol{W}_{12,t} = \left(W_{12,t}^{(1)}, W_{12,t}^{(2)}\right) \sim \text{BEV-logistic}(0.85),$$

$$\boldsymbol{W}_{13,t} = \left(W^{(1)}_{13,t}, W^{(3)}_{13,t}\right) \sim \text{BEV-logistic}(0.60),$$
$$\boldsymbol{W}_{23,t} = \left(W^{(2)}_{23,t}, W^{(3)}_{23,t}\right) \sim \text{BEV-logistic}(0.72).$$

Next, we generate observations from a trivariate extreme value (TEV) copula with logistic model and Fréchet margins, this time allowing the dependence parameter α_t to depend on atmosphere. That is, we have

$$\boldsymbol{W}_{123a,t} = \left(W^{(1)}_{123a,t}, W^{(2)}_{123a,t}, W^{(3)}_{123a,t}\right) \sim \text{TEV-logistic}(\alpha_t),$$

with $\alpha_t = (1 - |V_{8,t}|/6)$. Finally, we generate observations from a trivariate Gaussian copula with Fréchet margins; the pairwise correlations are set to $\rho_t = V^2_{8,t}/12$ and the resulting values are denoted

$$\boldsymbol{W}_{123b,t} = \left(W^{(1)}_{123b,t}, W^{(2)}_{123b,t}, W^{(3)}_{123b,t}\right).$$

The specific forms of α_t and ρ_t are chosen such that $\alpha_t, \rho_t \in (0, 1)$ for all given values of the atmosphere covariate, i.e., $\max_t |V_{8,t}| < \sqrt{12}$.

We then apply a max-mixture construction to the values generated from the five copulas listed above. For the first location, at time t, we set

$$W^*_{1,t} = \begin{cases} \max\left(W^{(1)}_{12,t}, W^{(1)}_{13,t}, W^{(1)}_{123a,t}\right)/3, & \text{if } V_{8,t} > 0, \\ \max\left(W^{(1)}_{12,t}, W^{(1)}_{13,t}, W^{(1)}_{123b,t}\right)/3, & \text{if } V_{8,t} \leq 0, \end{cases} \tag{5}$$

with $W^*_{2,t}$ and $W^*_{3,t}$ defined analogously. Finally, each $W^*_{i,t}$, $i = 1, 2, 3$, is transformed to Gumbel margins, with the resulting values corresponding to $Y_{i,t}$.

Again, we used Monte Carlo techniques to calculate the true values of p_1 and p_2. As part of the data challenge, teams were given $70 \times 300 = 21,000$ observations; we repeated this 20,000 times to obtain our final values for the truth, resulting in a total of 4.2×10^8 observations. We checked for convergence in our Monte Carlo estimates, to ensure that this size of simulation gave us sufficient accuracy to score and rank the teams. The true values we obtained were $p_1 = 5.38 \times 10^{-5}$ and $p_2 = 2.98 \times 10^{-5}$, to three significant figures.

4.3 Truth for C4

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_{50})$, with standard Gumbel distributed univariate margins. Furthermore, at time t the variable $\boldsymbol{Y}$ is denoted by $\boldsymbol{Y}_t$, and the set of vector random variables $\{\boldsymbol{Y}_t\}$ are independent and identically distributed over time t. We took $\boldsymbol{Y}$ to follow a clustered hierarchical model, with five clusters, with variables in different clusters (in the same cluster) being conditionally independent (conditionally dependent) of each other, given a 5-dimensional latent variable $\boldsymbol{Z}$. Only the ith component of $\boldsymbol{Z}_t$ has any effect on the behaviour of cluster i at time t. The vector random variables $\{\boldsymbol{Z}_t\}$ are

independent and identically distributed over time t with a standard marginal multivariate Gaussian distribution, with correlation matrix Σ with the off-diagonal entries being $\Sigma_{i,j} = 0.4$ $(i \neq j)$. As a consequence of this set up, the joint survivor function $\bar{F}_{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ is given by

$$\bar{F}_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_{\boldsymbol{z}\in\mathbb{R}^5} \left\{ \prod_{i=1}^{5} \bar{F}_i(\boldsymbol{y}_i; z_i) \right\} \phi_{\boldsymbol{Z}}(\boldsymbol{z}; \Sigma) d\boldsymbol{z}, \tag{6}$$

where $\bar{F}_i$ is the joint survivor function for cluster i, $\boldsymbol{y}_i$ corresponds to the ith cluster components of $\boldsymbol{y}$, $\phi_{\boldsymbol{Z}}$ is the multivariate normal joint density of $\boldsymbol{Z}$ and $\boldsymbol{z} = (z_1, \ldots, z_5)$. We additionally assume that each joint survivor function $\bar{F}_i$ is exchangeable, i.e., it is equal across any permutation of its arguments. The clustered variables are

$$\boldsymbol{C}_{1,t} = (Y_{1,t}, \ldots, Y_{8,t}), \qquad \boldsymbol{C}_{2,t} = (Y_{9,t}, \ldots, Y_{20,t}), \qquad \boldsymbol{C}_{3,t} = (Y_{21,t}, \ldots, Y_{33,t}), \tag{7}$$

$$\boldsymbol{C}_{4,t} = (Y_{34,t}, \ldots Y_{41,t}), \qquad \boldsymbol{C}_{5,t} = (Y_{42,t}, \ldots Y_{50,t}).$$

Here, $\boldsymbol{C}_{1,t}$ and $\boldsymbol{C}_{4,t}$ each have Gumbel distributed margins and a multivariate extreme value distribution copula with symmetric logistic form (Tawn, 1990), where the parameters are $0 < \alpha_t < 1$ and $0 < \gamma_t < 1$, respectively. For example, the form of the copula distribution function $C_{1,t}$ associated with $\boldsymbol{C}_{1,t}$ is

$$C_{1,t}(u_{1,t}, \ldots, u_{8,t}; \alpha_t) = \exp\left[- \left\{ \sum_{i=1}^{8} (-\log u_{i,t})^{1/\alpha_t} \right\}^{\alpha_t} \right],$$

with $u_{i,t} \in [0,1]$ for $i = 1, \ldots, 8$, for each $U_{i,t} \sim \text{Uniform}(0,1)$ random variable associated with the corresponding $Y_{i,t}$ in (7). Similarly, $\boldsymbol{C}_{2,t}$ and $\boldsymbol{C}_{5,t}$ each have Gumbel margins and an inverted multivariate extreme value distribution copula with symmetric logistic form (Ledford and Tawn, 1997), where the parameters are $0 < \beta_t < 1$ and $0 < \delta_t < 1$, respectively. For instance, for $\boldsymbol{C}_{2,t}$, the corresponding copula survivor function $\bar{C}_{2,t}$ has the form

$$\bar{C}_{2,t}(u_{9,t}, \ldots, u_{20,t}; \beta_t) = \exp\left(- \left[\sum_{i=9}^{20} \{-\log(1 - u_{i,t})\}^{1/\beta_t} \right]^{\beta_t} \right),$$

with $u_{i,t} \in [0,1]$ for $i = 9, \ldots, 20$, where each $U_{i,t} \sim \text{Uniform}(0,1)$ is associated with the corresponding $Y_{i,t}$ in (7). Cluster variables $\boldsymbol{C}_{3,t}$ have Gumbel margins and a Gaussian copula with dependence parameter $0 < \rho_t^c < 1$, where $\rho_t^c = 1 - \rho_t$, and ρ_t is the pairwise correlation coefficient parameter (in Gaussian margins) for all pairs of variables in $\boldsymbol{C}_{3,t}$.

Let $\boldsymbol{\theta}_t = (\theta_{1,t}, \ldots, \theta_{5,t}) = (\alpha_t, \beta_t, \rho_t^c, \gamma_t, \delta_t)$. We imposed that each $\theta_{i,t}$ would vary over t to avoid any team being able to recognise the within-cluster dependence structure, particularly as the core copulas we used are relatively standard, and feature widely as examples in our past research.

The parameter of each of the five copulas lies in the range $(0, 1)$ for all t, with dependence decreasing as each parameter increases, and with zero (one) corresponding to the vector variable having perfect dependence (independence) respectively; we imposed constraints on our parameter values to clearly avoid these boundary cases. In particular, we model the copula parameters at time t as functions of the latent variable $\boldsymbol{Z}_t$ via the relationship

$$\theta_{i,t}(z_{i,t}) = 0.4 + 0.5\Phi(z_{i,t}) \text{ for } i = 1, 2, 4, 5, \text{ and } \theta_{3,t}(z_{3,t}) = 0.1 + 0.6\Phi(z_{3,t})$$

where Φ is the standard univariate Gaussian distribution function and $z_{i,t}$ is a realisation of the ith component of $\boldsymbol{Z}_t$.

Between them, the five clusters exhibit both asymptotic dependence (AD) and asymptotic independence (AI), where we summarise the levels of each of these forms of extremal dependence through the multivariate versions of the measures χ and $\bar{\chi}$ introduced by Coles et al. (1999). We give the multivariate formulas for these quantities in Appendix A. Clusters 1 and 4 exhibit AD across all variables for all t with the coefficient of AD of the multivariate variable $\boldsymbol{C}_1$ being

$$\chi_{\boldsymbol{C}_1,t} = 8 - 28 \times 2^{\alpha_t} + 56 \times 3^{\alpha_t} - 70 \times 4^{\alpha_t} + 56 \times 5^{\alpha_t} - 28 \times 6^{\alpha_t} + 8 \times 7^{\alpha_t} - 8^{\alpha_t}$$

and $\chi_{\boldsymbol{C}_4,t}$ similarly with γ_t replacing α_t. In contrast, clusters 2, 3 and 5 all exhibit AI jointly across all variables, and for all pairs of variables, with bivariate $\bar{\chi}$ values being $\bar{\chi}_{\boldsymbol{C}_2,2,t} = 2^{1-\beta_t} - 1$, $\bar{\chi}_{\boldsymbol{C}_3,2,t} = \rho_t$, and $\bar{\chi}_{\boldsymbol{C}_5,2,t} = 2^{1-\delta_t} - 1$. The associated multivariate values of $\bar{\chi}$ are

$$\bar{\chi}_{\boldsymbol{C}_2,t} = [12^{1-\beta_t} - 1]/11,$$

$\bar{\chi}_{\boldsymbol{C}_3,t} = \rho_t$ (see Appendix B), and $\bar{\chi}_{\boldsymbol{C}_5,t} = [9^{1-\delta_t} - 1]/8$.

The final aspect of our specification is the allocation of the 50 sites to two sets, I_1 and I_2, corresponding to the regions U1 and U2, respectively. We did this allocation at random, subject to 25 sites being in each region and all clusters having at least one member in each region. When listed in the order as they appear in the simulated random vector $\boldsymbol{Y}$, the sets linking the components $\boldsymbol{Y}$ to the regions U_1 and U_2 are

$$\begin{aligned} I_1 &= \{5, 7, 8, 9, 11, 13, 16, 17, 19, 20, 31, 32, 33, 36, 37, 38, 39, 41, 44, 45, 46, 47, 48, 49, 50\} \\ I_2 &= \{1, 2, 3, 4, 6, 10, 12, 14, 15, 18, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 34, 35, 40, 42, 43\}. \end{aligned}$$

For the data challenge, the indices in I_1 and I_2 were not in ascending order, but instead a random permutation was applied, giving

$$\begin{aligned} I_1 &= \{46, 44, 37, 8, 50, 11, 47, 32, 45, 41, 17, 9, 19, 7, 39, 13, 49, 38, 5, 48, 33, 36, 20, 16, 31\} \\ I_2 &= \{22, 14, 2, 34, 3, 42, 27, 26, 21, 10, 12, 15, 6, 18, 23, 29, 30, 4, 1, 35, 43, 40, 28, 24, 25\}, \end{aligned}$$

which should be read as that the samples for Y_{46} were given as the data points for the first location of region U1 to the teams.

For computational evaluation of $\bar{F}_{\boldsymbol{Y}}(\boldsymbol{y})$, we used a Monte Carlo integration scheme for the five-dimensional integral in expression (6), i.e.,

$$\bar{F}_{\boldsymbol{Y}}(\boldsymbol{y}) \approx \sum_{j=1}^{m} \left[\prod_{i=1}^{5} \bar{F}_i \left\{ \boldsymbol{y}_i; \theta_i(z^*_{i,j}) \right\} \right] / m,$$

where we used the exact values for the cluster survivor functions and $z^*_{i,j}$ is the ith component of a simulated value of the random variable $\boldsymbol{Z}_j$. We took $m = 250,000$ to get the required level of accuracy in the estimate. To evaluate the joint cluster survivor functions we exploited the `R` functions `pmvevd` in the package `evd` (Stephenson, 2002) and `pmvnorm` in the package `mvtnorm` (Genz et al., 2021), as closed form expressions are not simple when $\phi_1 \neq \phi_2$.

To give some insight into the formulations, here we present the expressions for $\bar{F}_i$ when $\boldsymbol{y} = y\mathbf{1}$, such that $\Pr(Y_1 > y) = \phi$, i.e., as required for the evaluation of p_2. We now drop the t subscript from all notation, given that the data are independent and identically distributed over time. Specifically, with $\psi = 1 - \phi$, and as $|\boldsymbol{C}_1| = 8$,

$$\bar{F}_1(y\mathbf{1}; \alpha) = 1 - 8\psi + 28\psi^{2^\alpha} - 56\psi^{3^\alpha} + 70\psi^{4^\alpha} - 56\psi^{5^\alpha} + 28\psi^{6^\alpha} - 8\psi^{7^\alpha} + \psi^{8^\alpha},$$

with the result derived using the inclusion-exclusion formula from the simple expression for the joint distribution function. The expression is similar for cluster 4, as $|\boldsymbol{C}_4| = 8$. For cluster 2, $|\boldsymbol{C}_2| = 12$, resulting in $\bar{F}_2(y\mathbf{1}; \beta) = \phi^{12^\beta}$ and similarly for cluster 5 with $|\boldsymbol{C}_5| = 9$. Finally, for cluster 3, with $|\boldsymbol{C}_3| = 13$, the simplest evaluation is via the property $\bar{F}_3(y\mathbf{1}; \rho^c) = \bar{\Phi}_{13}(y_N\mathbf{1}; \mid \Sigma_\rho)$, where $\bar{\Phi}_{13}$ is a 13-dimensional joint survivor function of the standardised multivariate normal variable, $y_N = \Phi^{-1}(1 - \phi)$, and Σ_ρ is a correlation matrix with all off-diagonal entries being $\rho = 1 - \rho^c$.

The true values we obtained through the approach described above were $p_1 = 8.4 \times 10^{-23}$ and $p_2 = 5.4 \times 10^{-25}$, to two significant figures.

5 Performances of Teams

The rankings for each sub-challenge C1-C4, the cumulative points achieved and overall ranking are given for the teams by order of performance in Table 1. Congratulations to all teams, with every team achieving a top three finish in at least one of the four sub-challenges. Given this, all seven teams were invited to submit their methodology, results and subsequent reflections for review in the *Extremes* journal. We invited the top four teams in Table 1 to present two of the sub-challenges each at the EVA2023 conference in Milan. Specifically, Yahabe on C1 and C2, genEVA presented on C1 and C2, SHSmultiscale on C1 and C4, and Yalla on C3 and C4. Particular congratulations and recognition goes to Yalla from KAUST, who performed with excellence across all four sub-challenges. They were awarded a certificate at the conference dinner.

We should explain a little more about how the overall ranks were achieved. Our methods were set out in the initial announcement of the data challenge. Specifically, we identified that in the event of an overall tie in points between teams, the team with the better ranking in C4 would be overall ranked higher. Given the close nature of the competition between all of the teams after Yalla, it was not surprising that this

Place	Team	C1	C2	C3	C4	Points
1	Yalla	3	2	2	1	40
2	SHSmultiscale	1	6	7	2	31
3	genEVA	5	1	4	5	31
4	Yahabe	2	3	6	4	30
5	Uniofbathtopia	7	4	1	6	28
6	Lancopula Utopiversity	6	5	5	3	25
7	Wee_Extremes	4	7	3	7	23

Table 1 Final rankings of the competing teams, showing the ranking for each of the four sub-challenges and the overall points total.

tie-break rule was required to separate teams placed 2nd and 3rd. Other split decision rules were set out for sub-challenge ties.

The conversion of sub-challenge rankings to overall points also needs clarification. For each sub-challenge, we converted ranks into a points score according to the method used by the Eurovision Song Contest, i.e., 1st is 12 points, 2nd is 10, 3rd is 8, and then points decaying linearly (in steps of 1) with rank. We summed the points from the four sub-challenges to give the overall points for each team.

Of course, ranking can hide large and small differences between teams on individual challenges. It also does not reveal the improvement of skill levels over the period of the data challenge. In particular, at a mid-point we ranked the teams that entered their current best estimates for each sub-challenge and reported back to them only their ranking on each of the four sub-challenges. So, next we look at the estimates provided by the teams at the mid-point and end of the challenge to better understand the outcomes.

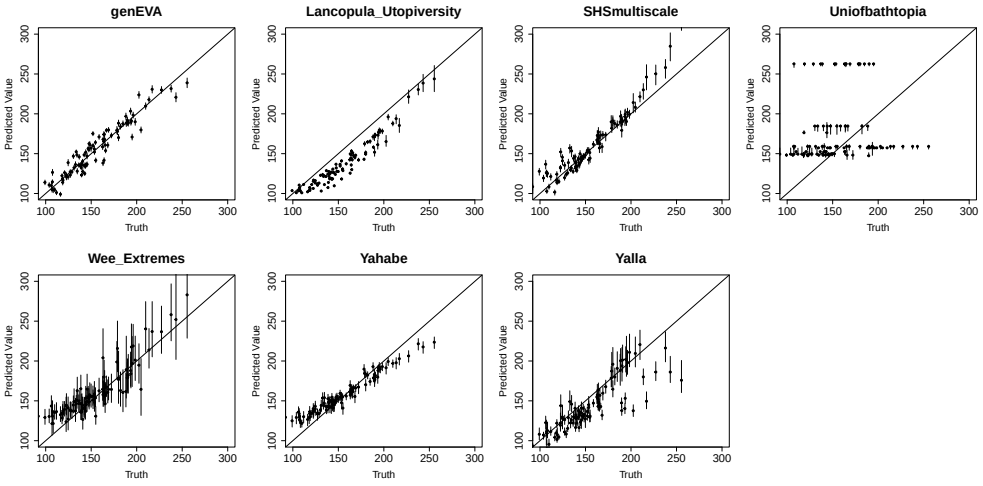

Fig. 1 Results for C1: for each team separately the estimated conditional quantile values $\{\hat{q}(\boldsymbol{x}_i) : i = 1, \ldots, 100\}$ against the corresponding true values $\{q(\boldsymbol{x}_i) : i = 1, \ldots, 100\}$, with the line of equality shown. Vertical bars show the associated 50% confidence intervals for each of the estimates.

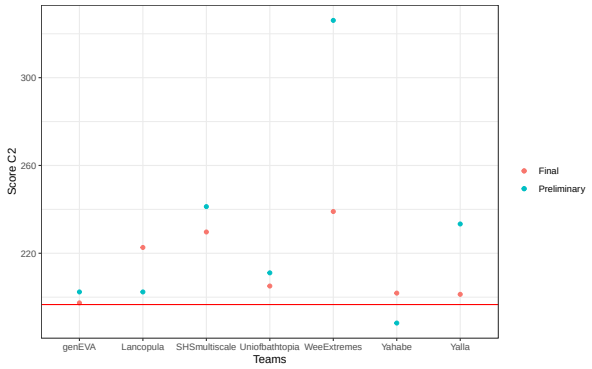

Fig. 2 Results for C2: showing estimated quantiles $\hat{q}$ both mid-point and final and the true q (solid line) across teams. Teams are ordered alphabetically along the x-axis.

For sub-challenge C1, the 100 intervals provided by the teams at the final stage of the challenge covered the true quantile values the following number of times: genEVA (30), Lancopula Utopiversity (6), SHSMultiscale (41), Wee_Extremes (64), Yahabe (36), Yalla (36) and Uniofbathtopia (3). Other than Wee_Extremes, all teams produced under-estimated intended coverage, though with the exception of two teams the coverages are reasonable approximations to the nominal level. To give greater insight into these performances, in Figure 1 we present a comparison between estimated conditional quantiles $\hat{q}(\boldsymbol{x}_i)$ and the associated true quantiles $q(\boldsymbol{x}_i)$ for $i = 1, \ldots, 100$; this plot also shows the associated 50% confidence intervals for each of the 100 conditional quantile estimates. This figure immediately explains the coverage problems for the two lowest-ranked teams, given the bias in their quantile estimates, with the other teams tending to perform well across the majority of estimates. The team best centred on the truth are genEVA, but they have very narrow confidence intervals, which is presumably why they slightly under-estimate coverage. Yalla mostly have estimates well centred on the truth, with exceptions for a few of the larger true values (indicating a key covariate may be missing in their model), but with wider confidence intervals the coverage is good. SHSmultiscale and Yahabe have their weakest performance when estimating the highest quantiles, with Wee_Extremes performing less well for lower (than higher) quantiles and having large confidence intervals, leading to over-coverage.

Figure 2 shows the teams' estimates of the marginal $(1 - (6 \times 10^4)^{-1})$th quantile for Y at Amaurot, both at the mid-point review and at the final submission. Recall that scoring is based on the loss function (1) that penalises errors in $\hat{q}$ below the true q more than an equal size error in $\hat{q}$ above q. Although team Yahabe's mid-point estimate was below q, their final submission gave an estimate above q, whilst all other teams' estimates were above q at both assessment times; hence, the final ranking was based on how close each team's estimate was to the true q. The top four teams were all very close on this sub-challenge. Given the analysis in C1, at first it is not too surprising that genEVA and Yalla did very well here as well. However, that outcome did not actually necessarily follow from the C1 analysis as it appears that most teams tackled C2 as a purely univariate series in Y rather than by marginalising out the covariates, as in Eastoe and Tawn (2009) or Rohrbeck et al. (2018), say.

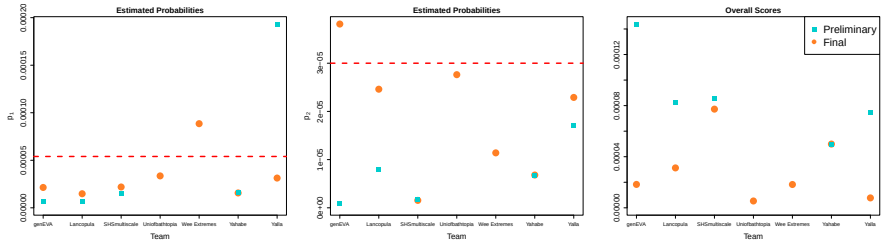

Fig. 3 Results for C3: $\hat{p}_1$ (left) $\hat{p}_2$ (centre) and overall score P_{12} (right). In the left and centre plots the true value is shown by the dashed line, in the right plot the best values are the lowest P_{12} scores. Teams are ordered alphabetically along the x-axis.

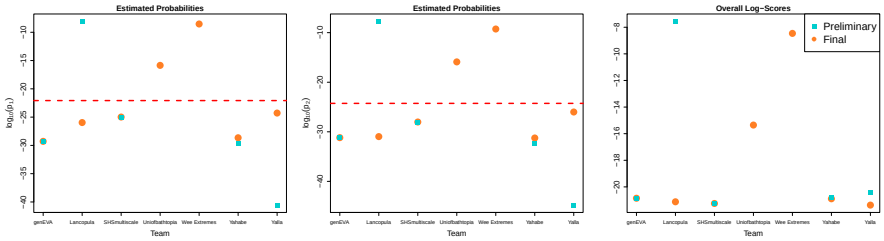

Fig. 4 Results for C4: $\log(\hat{p}_1)$ (left) $\log(\hat{p}_2)$ (centre) and overall log-score $\log(P_{12})$ (right). In the left and centre plots the true value is shown by the dashed line, in the right plot the best values are the lowest P_{12} scores. Teams are ordered alphabetically along the x-axis.

We now turn our attention to the multivariate challenges. Both Figures 3 and 4 have an identical structure, showing estimates $\hat{p}_1$ and $\hat{p}_2$ and scoring metric P_{12}, for C3 and C4, respectively. For C3, the teams typically under-estimated the two probabilities. For C4, the true probabilities are very small, probably smaller than any probability anyone has been set to estimate using extreme value methods! Despite the true values being so small in C4, the final performance across teams is very good, with the collective set of log-probability estimates, i.e., $\log(\hat{p}_i)$, centred on the true values. For each of C3 and C4, no team's performance deteriorated between the assessment times, but two teams only submitted their estimates at the final assessment round. For C3 and C4, the ranking was achieved based on the metric (2), with the smallest values for P_{12} being best. Here, we see that for C3 the top two teams were exceptional, whereas the five top teams were very close for C4.

6 Overview of the different approaches taken

6.1 Challenge C1

This challenge required point estimates for the conditional quantiles that exhibited low bias, in addition to confidence intervals that sufficiently represented the uncertainty in these point estimates and provided appropriate coverage. As seen from Figure 1, different teams performed well on the separate elements of this task, but to score well

overall a good performance was required from both perspectives. The top team, SHSmultiscale, performed very well across almost the entire range of the true conditional quantile values.

For obtaining point estimates of the conditional quantiles, all the teams started from a similar perspective by taking a GPD model for threshold exceedances and a threshold exceedance rate model, both of which could depend on covariates. Given that the starting point for all teams was so similar, it is very interesting to see such different performances being achieved. This is very revealing in terms of the level of subjectivity with respect to implementer choice when applying what are core tools in extreme value analysis.

The majority of teams used generalised additive models (GAMs) to allow smooth, but not fully parametric, variation in how the parameters of the tail models changed with covariates. The threshold choices varied from taking it as a constant to allowing it to change smoothly, dependent on covariates. Exceptions to this were the approach of genEVA (5th in this task), who used random forests for threshold selection and neural network formulations for both the scale and shape parameters, and the approach of Uniofbathtopia (7th in the task) who pooled response variation over covariates which were judged to have come from the same cluster - as Figure 1 shows, the latter approach placed too much restriction on the possible variation in quantile estimates over the different covariates.

There were important differences in the implementation of model fitting and checking. In particular, the top two teams in C1 (in terms of having the lowest bias in the point estimates) were the only teams using scoring rules; SHSmultiscale used the method of Gandy et al. (2022) and Yahabe used an interval scoring method of Gneiting and Raftery (2007). Wee_Extremes (4th on C1), with low bias in estimates, addressed the complexity of modelling in the GAM-GPD framework by performing model averaging.

The problem we set had the added challenge of 11.7% of covariate values being missing at random. The majority of teams simply discarded the entire data vector if any element was missing, which was a reasonable approach here given the relatively small proportion of the missingness and the large sample size of the data. However, Yahabe, Yalla and genEVA all used some form of imputation, and the variety of approaches they considered is interesting to contrast.

A key part of this challenge was to produce confidence intervals for the estimated conditional quantiles. Bootstrapping methods were at the core of the methods used: SHSmultiscale, Yalla, and Wee_Extremes used nonparametric approaches; while genEVA and Lancopula Utopiversity used semi-parametric approaches. In contrast, Yahabe used posterior predictive intervals. It appears that making more assumptions in the bootstrapping, via taking a semi-parametric rather than full non-parametric approach, has led to uniformly narrower intervals. These narrower intervals appear to be too short, as this has led to under-estimation of the intended coverage probabilities. This indicates that when using a form of semi-parametric confidence interval, one really needs to be sure that the truth is very close to the fitted model. We chose a particularly complex truth, so it was unlikely that any team's model would be sufficiently close to be able to rely on semi-parametric bootstrapping.

6.2 Challenge C2

This challenge required teams to optimise the choice of $\hat{q}$ such that it minimised the expected loss, i.e., to find $\hat{q}$ such that

$$\hat{q} = \text{argmin}_{y \in \mathbb{R}} L(q, y), \tag{8}$$

where L is defined by expression (1) and q is given by $\Pr(Y > q) = (6 \times 10^4)^{-1}$, but q is unknown as the distribution of Y is unknown. So the core effort in this problem is to estimate the marginal distribution of Y, far into its upper tail, or equivalently to obtain an estimate of the required value of q, before performing optimisation (8). We were pleased to see that almost all teams got to grips well with the non-standard extreme value problem of working with an asymmetric loss function, penalising under-estimation for their quantile estimation. There were also different ways that the loss function was incorporated into teams' solutions, e.g., SHSmultiscale embedded this into a measure of fit within their use of methods from Stupfler and Usseglio-Carleve (2021).

As this challenge was explicitly about the marginal distribution of Y, and we did not directly state that the use of the conditional model was required, it is maybe not too surprising that all but team Yahabe (3rd for this challenge) took the approach to model the distribution of Y without reference to the effects of covariates. In fact, Yahabe also tried a purely marginal approach but found it to perform less well than using their C1 models as a basis for their proposed answer. Teams other than Yahabe analysed the observed marginal sample of Y data, fitting models motivated by univariate extreme value theory. The two top teams, genEVA and Yalla, were very successful with their approaches, which combined classical extreme value methods with modern statistical/machine learning techniques. For example, Yalla used used their conditional model from C1 to generate data for their training set for a neural Bayes estimator.

If we were given this challenge, we expect that our approach, like Yahabe, would instead have been to use the information from the covariates, particularly given the substantial modelling effort that the teams had already invested in challenge C1. Specifically, we would have estimated the marginal for Y by exploiting the formulation

$$\Pr(Y > y) = \int_{\boldsymbol{x} \in \mathcal{X}} \Pr(Y > y \mid \boldsymbol{X} = \boldsymbol{x}) dF_{\boldsymbol{X}}(\boldsymbol{x}),$$

where $\mathcal{X}$ and $F_{\boldsymbol{X}}$ are the domain and the joint distribution respectively, of the covariates $\boldsymbol{X}$. As $F_{\boldsymbol{X}}$ is unknown, we had expected the teams to use the empirical joint distribution of the sample of $N = 21,000$ realisations from $\boldsymbol{X}$ in the observed data. The empirical estimator should be reliable, as the C1 set up pointed to the periodic behaviour in the covariates over the span of the observations. Using this empirical estimator leads to the estimated marginal survivor function of Y being

$$\Pr(Y > y) = \frac{1}{N} \sum_{i=1}^{N} \Pr(Y > y \mid \boldsymbol{X} = \boldsymbol{x}_i), \tag{9}$$

where $\{\boldsymbol{x}_i : i = 1, \ldots, N\}$ is the sample of covariate realisations. Then, using the estimator of $\Pr(Y > y \mid \boldsymbol{X} = \boldsymbol{x})$ from C1 in expression (9), the value of q can be used in expression (8) to estimate $\hat{q}$.

We view that there is a drawback of using a univariate approach in this problem, as you lose all information about the covariates, despite knowing a lot about their distribution from the large observed sample and from the contextual information provided in the challenge brief. Ignoring the covariates leaves the analysis at the mercy of the specifics of the extremes in the observed random sample of Y, which we feel can be avoided through our suggested conditional approach. Clearly, the univariate approach requires much less modelling effort than the conditional approach, and if the fit of the univariate model seems sufficient in the tail, that added simplicity is highly appealing.

6.3 Challenge C3

It was particularly interesting to see that this challenge was addressed by a wide range of different methods across the teams. Despite the variation in strategies, it is notable from Figure 3 that almost all teams under-estimated p_1 and p_2 for this challenge.

The most successful team in this challenge was Uniofbathtopia. They adopted a modelling approach based on multivariate max-linear models (Fougères et al., 2013), sparse projections (Meyer and Wintenberger, 2024) and the estimation procedure for joint tail probabilities by Kiriliouk and Zhou (2022). Such an approach works ideally for this problem, as the true generating process for the data covered a mixture of AD (over the triple) and at least one component of the triple being independent of the other components which were AD (pairwise). In particular, the max-linear model handled the mixture structure very well. Furthermore, the sparse projections ensure that the AI terms are likely to be correctly identified as placing mass on the associated boundaries of the spectral measure, i.e., the edge faces of the triangular simplex.

Yalla, Yahabe and SHSmultiscale (2nd, 6th and 7th on this challenge, respectively) all used the conditional multivariate extreme value models stemming from Heffernan and Tawn (2004), which allows the dependence in the AI components of the multivariate variable to be modelled, not simply to be treated as independent and AD as by Uniofbathtopia. So Yalla's approach is very successful as it performs well without making such strict (but correct here) assumptions as made by Uniofbathtopia. However, Yahabe's and SHSmultiscale's less strong placing suggests that the way the method was implemented may be important. It was interesting to see that Yahabe had implemented a wide range of models but their benchmark testing at less extreme levels led to their choice of a conditional multivariate extreme value model as the basis for their submitted challenge entry.

The teams in 3rd to 5th places all performed rather similarly in their overall score, despite their very different approaches: Wee_Extremes taking an off-the-shelf vine copula approach (Dissmann et al., 2013), in no way tuned for extreme value problems; genEVA either breaking down the problems into univariate formulations to avoid having to fully model dependence (for p_2), or using a Gaussian copula model (for p_1); and Lancopula Utopiversity extending the joint survivor function asymptotic approach of Murphy-Barltrop and Wadsworth (2024), which allows for both AD and AI. Yahabe

were extensive in their analysis, applying a wide range of methods including developing an extension of the conditional approach of Heffernan and Tawn (2004), which can also capture both AD and AI, to allow for a fully parametric joint model for the residuals in this structure.

A number of teams explored using the suggested covariates in their analysis. Some teams found no clear cut effect of the covariates, so they opted for a simpler approach of assuming all vector variables to be independent and identically distributed. This finding indicates that these covariates, though present, were not very informative for this challenge. This interpretation is further supported by the fact that the teams who excluded covariates from their models performed well. Although Lancopula Utopiversity, SHSmultiscale and Yalla all include the suggested covariates in the parameters of their models. There were also a range of very helpful exploratory analyses. For example, we particularly liked SHSmultiscale's assessment of the probability values relative to cases of all components in the triple exhibiting perfect dependence and independence, and the use by genEVA and Lancopula Utopiversity of pairwise estimates of the measures of AD and AI, χ and $\bar{\chi}$ respectively (Coles et al., 1999), to assess whether the given covariate influenced the extreme values of the variables of interest.

6.4 Challenge C4

With the exception of Yalla who did consistently very well on both C3 and C4, it is worth noting that the teams who did particularly well (less well) on C3 tended to do worse (better) on C4. There does not appear to be any particular reason for this outcome, but perhaps the more simplified methods that worked well in the straightforward context of C3, were exposed in their application to the higher dimensional challenge of C4, which required explicit use of methods tailored to describing the sub-asymptotic dependence between AI variables. We found the teams' performances to be highly impressive, as the complex dependence structure we used to generate the data does not arise in any previously published multivariate extreme value problems. In particular, we had clusters of sites which had different forms of extremal dependence (AD or AI) and different levels of dependence within each of these clusters.

It was pleasing that almost all teams correctly identified the clustering structure that we used in the simulated data generation. To help in this identification of clusters, the teams used rather similar pairwise estimates of rank-based correlation measures and the χ and $\bar{\chi}$ measures of AD and AI respectively (Coles et al., 1999) (or variants of these such as the extremal variogram (Davis and Mikosch, 2009), the F-madogram (Guillou et al., 2014), and extremal dependence measure of (Larsson and Resnick, 2012)) and methods from factor analysis. From these exploratory analyses, the teams seemed to pick out the cluster structure well. Although the cluster structure was correctly identified, Yahabe were unique in investigating evidence for, and correctly finding, that the dependence within each of these clusters was exchangeable, for which they should be highly commended.

After identifying the clusters, the teams assumed independence between variables in different clusters which, as seen from Section 4.3, was incorrect. However, the dependence between variables from different clusters is very weak, coming only through our hierarchical model, so the parameters of the copulas for different clusters are dependent

but the observations are conditionally independent given these parameters. Consequently, we do not believe the team's incorrect assumption of independence between clusters was very influential in their performance.

As with C3, again the teams used a range of methods. As Figure 4 shows, five of the teams performed very well, relative to the other two teams. So our discussion will focus on the top group, covering them in ranking order from the top. Here we focus on their within-cluster dependence modelling. Yalla used a non-parametric estimator proposed by Krupskii and Joe (2019) for estimation of joint exceedance probabilities; SHSmultiscale and Lancopula Utopiversity applied the conditional multivariate extremes model of Heffernan and Tawn (2004); Yahabe investigated whether clusters could be classified as AD or AI, and for the former fitted parametric models and for the latter they used the semi-parametric methods of Heffernan and Tawn (2004); and genEVA used empirical estimates for clusters they identified as AD, and exploited factorisations into univariate probabilities for the AI clusters which could each be modelled using a GPD.

This challenge was particularly difficult as the true probabilities were so very small, given the very weak dependence between clusters, and the fact that there was a wide variety of different AD or AI dependence forms within each cluster. The issue of dealing with a 50-dimensional problem, and clusters up to 12-dimensions, was probably outside of the experience range of all extreme value analysts, so we would really like to praise all the teams for their efforts on this challenge.

Declarations

- Funding: No funding to be reported
- Competing interests: The authors declare that they have no conflict of interest
- Availability of data: The datasets given to the teams for the EVA data challenge are freely available via the University of Bath Research Data Archive under doi.org/10.15125/BATH-01399.

Appendix A Multivariate extremes measures of AD and AI

Let $D = \{1, \ldots, d\}$ be the indices of d marginally identically distributed random variables $(Y_1, \ldots, Y_d)$, with upper marginal endpoint y^F. Here, we define the natural extensions of the bivariate measures χ of AD and $\bar{\chi}$ of AI given by Coles et al. (1999) to give equivalent versions of these two tail indices for multivariate variables, i.e., $(Y_1, \ldots, Y_d)$ with $d \geq 2$, with the expressions being identical to those in Coles et al. (1999) when $d = 2$. Specifically, the coefficient of AD of a multivariate variable is given by

$$\chi_D = \lim_{y \to y^F} \Pr(Y_2 > y, \ldots, Y_d > y \mid Y_1 > y),$$

and the coefficient of AI of a multivariate variable is $\bar{\chi}_D = (d\eta_D - 1)/(d - 1)$, where η_D is defined implicitly through the expression

$$\Pr(Y_1 > y, \ldots, Y_d > y) = \mathcal{L}_D \left\{1/\Pr(Y_1 > y)\right\} \left\{\Pr(Y_1 > y)\right\}^{1/\eta_D}, \qquad \text{(A1)}$$

as $y \to y^F$, with $\mathcal{L}_D$ a slowly varying function at infinity. The η_D term in the power decay of the joint probability that all components of the vector variable are large has the property $0 < \eta_D \leq 1$, so it follows that $-1/(d-1) < \bar{\chi}_D \leq 1$. Both χ_D and η_D have been considered previously by Eastoe and Tawn (2012), with larger values of χ_D (and $\bar{\chi}_D$) indicating stronger levels of AD (and AI) respectively.

If $(Y_1, \ldots, Y_d)$ are AD (AI), then $\chi_D > 0$ ($\chi_D = 0$) respectively, with the value of χ_D measuring the degree of AD. However, $\chi_D = 0$ does not imply that the lower dimensional joint distributions of the variables indexed by D are not AD, that issue would need to be separately considered by the evaluation of χ_B, for the appropriate $B \subset D$. If $\bar{\chi}_D = 1$ and $\mathcal{L}_D(y) \not\to 0$ as $y \to \infty$ then $(Y_1, \ldots, Y_d)$ are AD otherwise they are AI with the degree of AI given by $\bar{\chi}_D$. When all variables are mutually independent, $\bar{\chi}_D = 0$, and when they are all perfectly dependent, $\chi_D = 1$. Positive extremal dependence corresponds to $\max\{\chi_D, \bar{\chi}_D\} > 0$, with negative extremal dependence giving $\bar{\chi}_D < 0$.

Appendix B Value of $\bar{\chi}_D$ for a Gaussian copula with equal pairwise correlations

For one of the clusters in challenge C4, we are interested in a d-dimensional Gaussian copula with covariance matrix Σ having elements $\sigma_{i,i} = 1$, $i = 1, \ldots, d$, and all $\sigma_{i,j} = \rho \in [0, 1)$, $i \neq j$. From results in Nolde (2014) and Nolde and Wadsworth (2022), it can be deduced that in this case, η_D, defined in (A1), is equal to $|\Sigma^{-1}|^{-1}$, where $|\cdot|$ is the sum of all the elements of the matrix. We now show that this corresponds to $\bar{\chi}_D = \rho$.

First, note that

$$\Sigma = (1-\rho) I_d + \rho \mathbf{1}\mathbf{1}^T,$$

where I_d is the $d \times d$ identity matrix and $\mathbf{1}$ is a column vector of length d where all elements are one. Then, applying the Sherman-Morrison formula (see, e.g., Bartlett (1951)) gives the inverse of the correlation matrix as

$$\Sigma^{-1} = (1-\rho)^{-1} I_d - \left\{ \frac{\rho}{(1-\rho)^2 + d(1-\rho)\rho} \right\} \mathbf{1}\mathbf{1}^T,$$

from which we have

$$\begin{aligned} |\Sigma^{-1}| &= d(1-\rho)^{-1} - d^2 \left\{ \frac{\rho}{(1-\rho)^2 + d(1-\rho)\rho} \right\} \\ &= \frac{d(1-\rho) + d^2\rho}{(1-\rho)^2 + d(1-\rho)\rho} - \frac{d^2\rho}{(1-\rho)^2 + d(1-\rho)\rho} \\ &= \frac{d}{1 + (d-1)\rho}. \end{aligned}$$

This yields $\eta_D = \{1 + (d-1)\rho\}/d$, and substituting this into the equation for $\bar{\chi}_D$ in Appendix A, we have $\bar{\chi}_D = (d\eta_D - 1)/(d-1) = \rho$.

References

Smith, R.L.: Bayesian and frequentist approaches to parametric predictive inference (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics 6, pp. 589–612. Oxford University Press, London (1999)

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

Gumbel, E.J.: Bivariate exponential distributions. Journal of the American Statistical Association **55**(292), 698–707 (1960)

Tawn, J.A.: Modelling multivariate extreme value distributions. Biometrika **77**(2), 245–253 (1990)

Ledford, A.W., Tawn, J.A.: Modelling dependence within joint tail regions. Journal of the Royal Statistical Society: Series B (Methodological) **59**(2), 475–499 (1997)

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

Stephenson, A.G.: evd: Extreme value distributions. R News **2**(2), 31–32 (2002)

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T.: mvtnorm: Multivariate Normal and T Distributions. (2021). R package version 1.1-2. https://CRAN.R-project.org/package=mvtnorm

Eastoe, E.F., Tawn, J.A.: Modelling non-stationary extremes with application to surface level ozone. Journal of the Royal Statistical Society: Series C (Applied Statistics) **58**(1), 25–45 (2009)

Rohrbeck, C., Eastoe, E.F., Frigessi, A., Tawn, J.A.: Extreme value modelling of water-related insurance claims. The Annals of Applied Statistics **12**(1), 246–282 (2018)

Gandy, A., Jana, K., Veraart, A.E.: Scoring predictions at extreme quantiles. Advances in Statistical Analysis **106**(4), 527–544 (2022)

Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**(477), 359–378 (2007)

Stupfler, G., Usseglio-Carleve, A.: Composite bias-reduced L^p-quantile-based estimators of extreme quantiles and expectiles. Canadian Journal of Statistics **51**(2), 704–742 (2021)

Fougères, A.-L., Mercadier, C., Nolan, J.P.: Dense classes of multivariate extreme value distributions. Journal of Multivariate Analysis **116**, 109–129 (2013)

Meyer, N., Wintenberger, O.: Multivariate sparse clustering for extremes. Journal of the American Statistical Association **119**, 1911–1922 (2024)

Kiriliouk, A., Zhou, C.: Estimating probabilities of multivariate failure sets based on pairwise tail dependence coefficients. arXiv preprint arXiv:2210.12618 (2022)

Heffernan, J.E., Tawn, J.A.: A conditional approach for multivariate extreme values (with discussion). Journal of the Royal Statistical Society: Series B (Methodological) **66**(3), 1–34 (2004)

Dissmann, J., Brechmann, E.C., Czado, C., Kurowicka, D.: Selecting and estimating regular vine copulae and application to financial returns. Computational Statistics & Data Analysis **59**, 52–69 (2013)

Murphy-Barltrop, C.J.R., Wadsworth, J.L.: Modelling non-stationarity in asymptotically independent extremes. Computational Statistics & Data Analysis **1999**, 108025 (2024)

Davis, R.A., Mikosch, T.: The extremogram: A correlogram for extreme events. Bernoulli **15**(4), 977–1009 (2009)

Guillou, A., Naveau, P., Schorgen, A.: Madogram and asymptotic independence among maxima. REVSTAT - Statistical Journal **12**(2), 119–134 (2014)

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

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

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

Nolde, N.: Geometric interpretation of the residual dependence coefficient. Journal of Multivariate Analysis **123**, 85–95 (2014)

Nolde, N., Wadsworth, J.L.: Linking representations for multivariate extremes via a limit set. Advances in Applied Probability **54**(3), 688–717 (2022)

Bartlett, M.S.: An inverse matrix adjustment arising in discriminant analysis. The Annals of Mathematical Statistics **22**(1), 107–111 (1951)