Supplementary material for "Modelling sub-daily precipitation extremes with the blended generalised extreme value distribution"

Silius M. Vandeskog[1] Sara Martino[1] Daniela Castro-Camilo[2]
Håvard Rue[3]

[1] Norwegian University of Science and Technology (NTNU)
[2] University of Glasgow
[3] King Abdullah University of Science and Technology (KAUST)

S1 Simulation study

A simulation study is conducted for comparing the performance of the generalised extreme value (GEV) distribution and blended generalised extreme value (bGEV) distribution in a univariate setting. Our penultimate goal when modelling block maxima is to estimate return levels. Thus, the two distributions are compared by evaluating the performance of their return level estimators.

We sample $n \in \{25, 50, 100, 500, 1000\}$ i.i.d. "block maxima" from a GEV distribution resembling the estimated distribution of the yearly maxima of hourly precipitation. From Table 3 and Table 4 in "Modelling sub-daily precipitation extremes with the blended generalised extreme value distribution" we find that the posterior mean of the intercept in $\boldsymbol{\beta}_\mu$ is $E[\beta_{\mu,0}] = 0.661$, and the posterior mean of the intercept in $\boldsymbol{\beta}_\sigma$ is $E[\beta_{\sigma,0}] = 2.835$. Consequently, we set $\alpha = 0.5$, $\beta = 0.8$, $\mu_\alpha = E[\beta_{\mu,0}] \cdot \exp(E[\beta_{\sigma,0}]) \approx 11.26$, and $\sigma_\beta = E[\sigma_\beta^*] \cdot \exp(E[\beta_{\sigma,0}]) \approx 2.01$. The tail parameter is set equal to the posterior mean from Table 4: $\xi = 0.178$. Using the n block maxima, we compute maximum likelihood estimators for the parameters of the GEV and bGEV distributions, which are further used for computing return level estimators for periods of length 25, 50, 100, 250 and 500. The actual maximum likelihood estimation is performed on the (μ, σ, ξ) parametrisation, in which $\mu \approx 10.05$ and $\sigma \approx 3.21$. Optimisation is performed using the `optim` function in `R`, where the true values of (μ, σ, ξ) are used as initial values. The entire procedure is repeated 500 times for each value of n. This allows us to estimate the distribution of the maximum likelihood estimators for all return levels in question.

Figure S1.1 displays the distribution of all return level estimators using both the GEV and the bGEV distributions. The difference between the distributions of the GEV estimators and the bGEV estimators are negligible for small values of n. For large return periods and values of n some slight differences are appearing. However, in practical situations one rarely encounters as much as 1000 block maxima, and the blocks are almost never large enough for the limit distribution to hold exactly. Thus, the differences in distribution can be considered negligible for large values of n for almost all real-life settings. Larger differences in model fit could theoretically occur for small blocks where the block maxima deviate further from

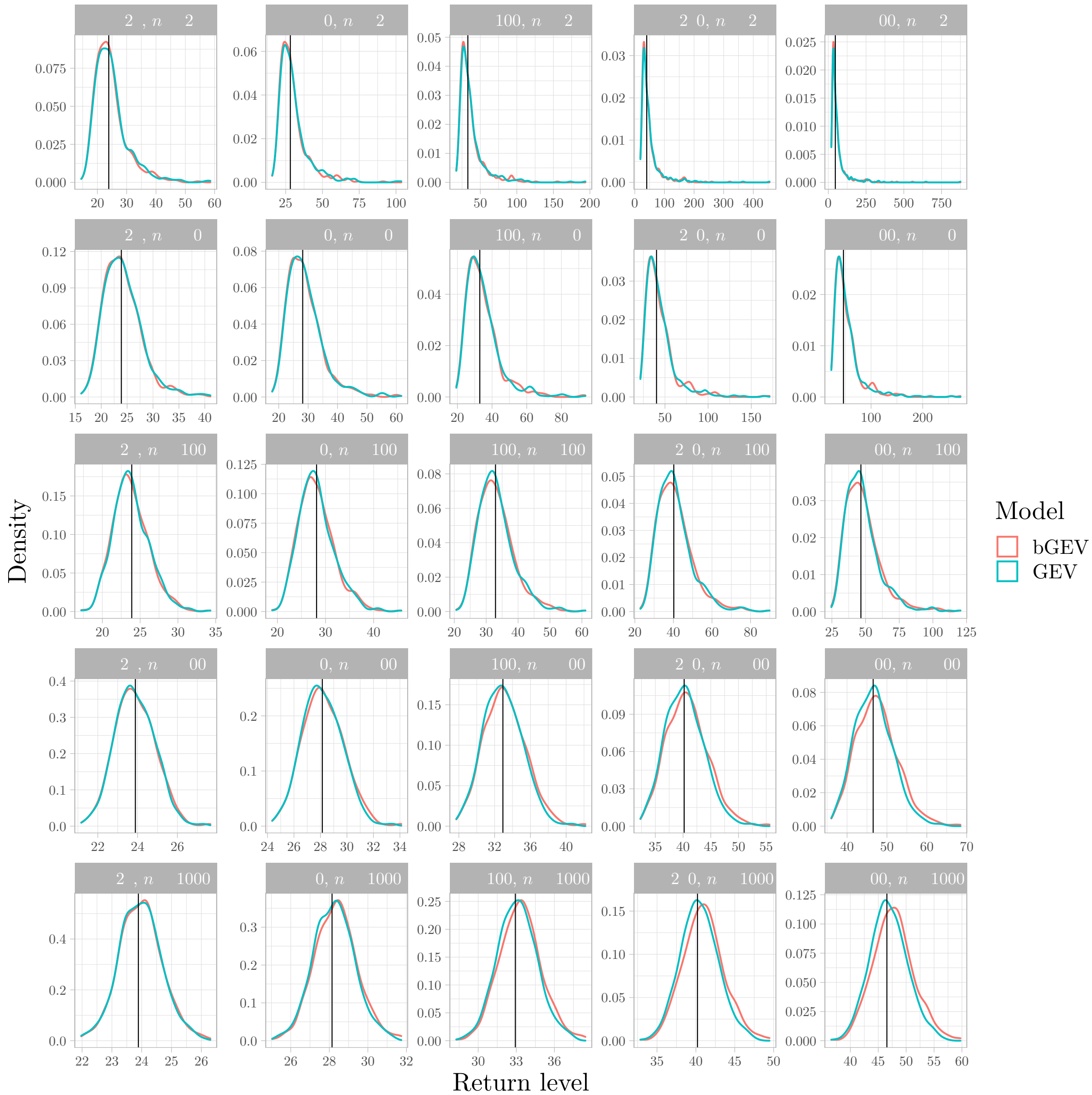

Figure S1.1: Distributions of return level estimators computed from model fits with the GEV and bGEV distributions to simulated data from a GEV distribution. The dark vertical lines display the true return levels. Return periods are denoted as T, with unit “block sizes”, and the number of block maxima used for fitting the GEV and bGEV distributions is denoted as n.

the GEV distribution. However, in such settings, both the GEV and the bGEV distribution would be misspecified, and we have no reason to believe that the misspecification in the GEV distribution would be strictly better or worse than the misspecification in the bGEV distribution. Therefore, our results imply that we do not lose anything in performance by modelling GEV data with the bGEV distribution instead of the GEV distribution.

In most real-life settings we do not have as much prior knowledge about the true underlying distribution, and we will likely choose less optimal initial values and/or prior distributions. To examine the effects of choosing less optimal initial values, we repeat the maximum likelihood estimation on the exact same data used for Figure S1.1, using different initial values. Figure S1.2 displays the distributions of the return level estimators when we use an initial value of 0.9 instead of 3.2 for σ, without changing the initial values for μ or ξ. For small values of n the two distributions yield comparable results. However, for large values of n, the performance of the GEV distribution quickly deteriorates, and the return level estimators are shifted to the left. The bGEV estimators, however, seem to have identical distributions in Figures S1.1 and S1.2, and are not affected by the slight change of initial values. This demonstrates how inference with the bGEV distribution is more robust than inference with the GEV distribution. The parameter-dependent support of the GEV complicates likelihood-based inference and makes it more difficult to locate the global maximum of the likelihood, even in simple and univariate settings with unrealistically large amounts of data that are perfectly GEV-distributed.

In such a simple setting, it is relatively easy to see that we chose bad initial values for (μ, σ, ξ). However, when modelling block maxima with a large number of covariates and in high-dimensional settings, it becomes considerably more difficult to choose good enough initial values and/or prior distributions. Additionally, it is not trivial to examine a maximum likelihood estimator or posterior distribution and find out if the inference procedure was successful or not. The GEV return level estimators in Figure S1.2 take on reasonable values and are large enough to seem realistic. Without knowing the truth, it would not be trivial for a practitioner to detect that they are faulty, and caused by numerical problems. Thus, if one performs likelihood-based inference using the GEV distribution, one should always perform extensive checks to evaluate whether the fit is reasonable, or if it is caused by numerical problems. The increased robustness of the bGEV distribution provides more safety, and most practitioners can be more confident in their model fits after fitting the bGEV distribution to block maxima data instead of the GEV distribution.

Similarly, while attempting to model return levels for precipitation in Norway using the `gev` family in `R-INLA`, we experienced that e.g. adding or removing one covariate from the model could completely alter the model fits if our priors were too wide or the data were not correctly standardised. Examining which of the fits were the best required comprehensive and time-demanding cross-validation studies, and it was difficult to find out if the best model fit was "good enough", or if a slight change in priors or covariates would yield considerably better model fits. These problems were considerably mitigated by switching to the `bgev` family, which resulted in more stable inference using less informative priors, and less optimal initial values.

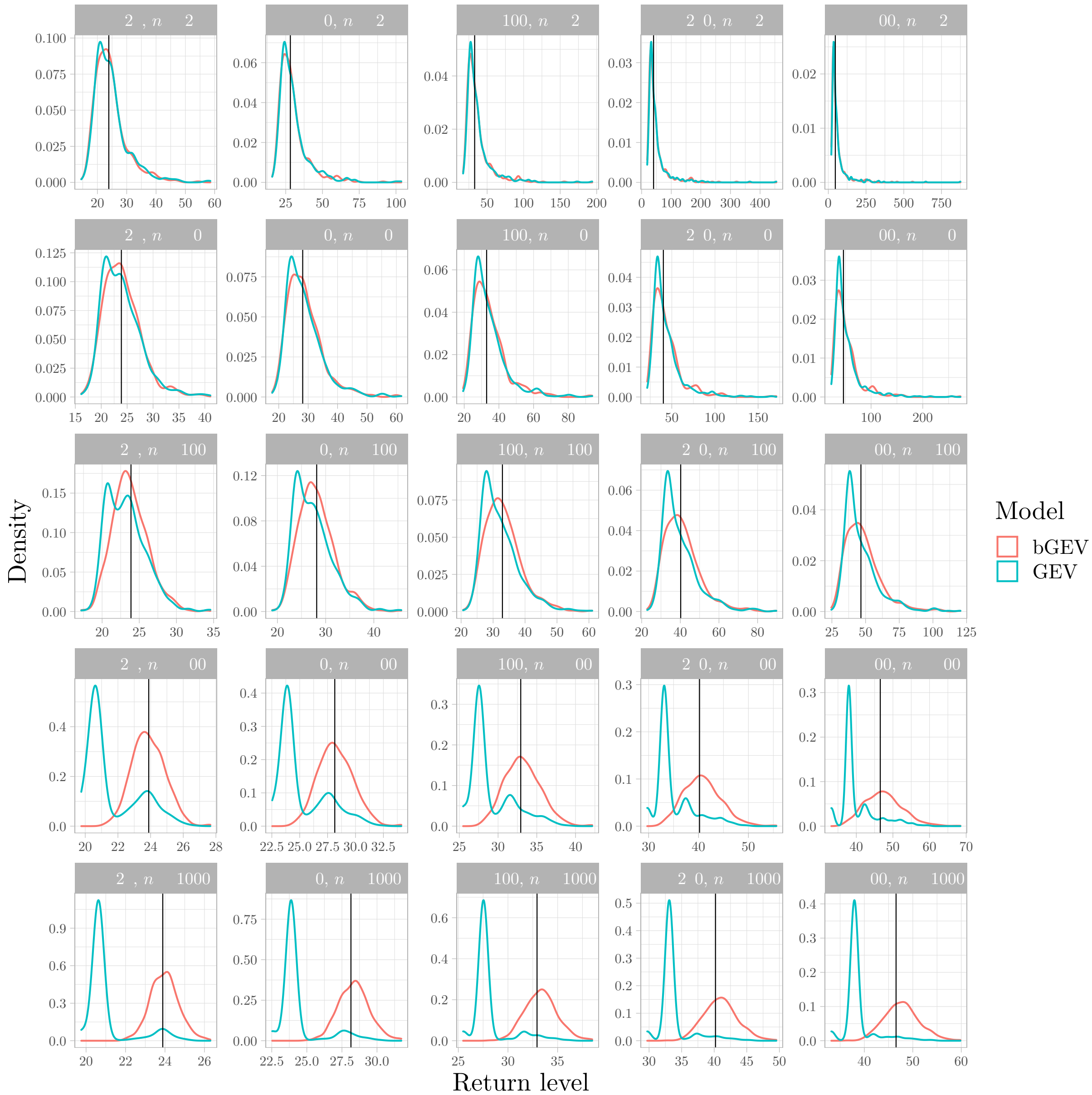

Figure S1.2: Distributions of return level estimators computed from model fits with the GEV and bGEV distributions to simulated data from a GEV distribution, when using bad initial values. The dark vertical lines display the true return levels. Return periods are denoted as T, with unit "block sizes", and the number of block maxima used for fitting the GEV and bGEV distributions is denoted as n.

S2 PC prior for the tail parameter

In order to compute the PC prior for the GEV and bGEV distribution with $\xi \geq 0$ and base model $\xi = 0$, one must first compute $\mathrm{KLD}(\pi_\xi, \pi_0)$ for the GEV distribution and the bGEV distribution. Notice that the base model is identical for both distributions. Writing the GEV distribution function as $F(x) = \exp(-h(x))$ gives the probability density function

$$\pi_\xi(x) = -\exp(-h(x))h'(x), \quad h(x) = (1 + \xi(x-\mu)/\sigma)_+^{-1/\xi}.$$

The KLD for the GEV distribution is equal to

$$\begin{aligned}
\mathrm{KLD}(\pi_\xi, \pi_0) &= \int \pi_\xi(x) \log\left(\frac{\pi_\xi(x)}{\pi_0(x)}\right) \mathrm{d}x \\
&= \int -e^{-h(x)} h'(x) \cdot \left(-h(x) + \log(-\sigma h'(x)) + \exp\left(-\frac{x-\mu}{\sigma}\right) + \frac{x-\mu}{\sigma}\right) \mathrm{d}x.
\end{aligned}$$

The fourth term of the KLD is simply equal to the expectation of $(x-\mu)/\sigma$, which is known. The first term is easily solvable with the substitution $u = h(x)$. Using the same substitution for the second term with the knowledge that $\log(-\sigma h'(x)) = (1+\xi)\log(h(x))$ gives the integral

$$\int -e^{-h(x)} h'(x) \log(-\sigma h'(x)) = \int_0^\infty e^{-u}(1+\xi)\log(u)\ \mathrm{d}u = -(1+\xi)\gamma,$$

where γ is the Euler-Mascheroni constant. We are unable to find a closed-form expression for the third term. However, using substitution with $u = h(x)$ gives

$$\int -e^{-h(x)} h'(x) \exp\left(-\frac{x-\mu}{\sigma}\right) \mathrm{d}x = \int_0^\infty \exp\left(\frac{1}{\xi} - u - \frac{u^{-\xi}}{\xi}\right) \mathrm{d}u,$$

which is finite for $\xi > 0$. With a change of variables, this integral is transformed to have finite limits,

$$\int_0^\infty \exp\left(\frac{1}{\xi} - \left(u + u^{-\xi}/\xi\right)\right) \mathrm{d}u = \int_0^1 \exp\left(\frac{1}{\xi}\left(1 - (-\log v)^{-\xi}\right)\right) \mathrm{d}v.$$

This can easily be numerically approximated. Summarising, the KLD for the GEV is finite for $0 \leq \xi < 1$ and equal to

$$\mathrm{KLD}(\pi_\xi, \pi_0) = -1 - (1+\xi)\gamma + \xi^{-1}(\Gamma(1-\xi) - 1) + \int_0^1 \exp\left(\frac{1}{\xi}\left(1 - (-\log v)^{-\xi}\right)\right) \mathrm{d}v. \quad \text{(S2.1)}$$

The PC prior has probability density function $\pi(d) = \lambda \exp(-\lambda d)$ with $d = \sqrt{2\mathrm{KLD}(\pi_\xi, \pi_0)}$. Transforming the PC prior from a distribution on d to a distribution on ξ gives

$$\pi(\xi) = \lambda \exp\left(-\lambda d(\xi)\right) \left|\frac{\partial d(\xi)}{\partial \xi}\right| = \frac{\lambda}{d(\xi)} \exp\left(-\lambda d(\xi)\right) \left|\frac{\partial}{\partial \xi}\mathrm{KLD}(\pi_\xi, \pi_0)\right|. \quad \text{(S2.2)}$$

Consequently, the derivative of the KLD must also be computed. Using derivation under the integral sign gives

$$\frac{\partial}{\partial \xi}\mathrm{KLD}(\pi_\xi, \pi_0) = -\gamma - \frac{\Gamma(1-\xi)\Psi(1-\xi)}{\xi} - \frac{\Gamma(1-\xi)-1}{\xi^2} + \int_0^1 g(v;\xi)\mathrm{d}v,$$

where $\Psi(\cdot)$ is the digamma function, and

$$g(v;\xi) = \exp\left(\frac{1}{\xi}\left(1-(-\log v)^{-\xi}\right)\right)\frac{1}{\xi^2}\left(-1+(-\log v)^{-\xi}\left(1+\xi\log(-\log v)\right)\right).$$

This integral must also be numerically approximated.

When computing the KLD for the bGEV distribution, the integral for the KLD is divided into three parts:

$$\begin{aligned}
&\mathrm{KLD}(\pi_\xi, \pi_0) = \\
&\quad \int_{-\infty}^{p_a} \pi_\xi(x)\log\left(\frac{\pi_\xi(x)}{\pi_0(x)}\right)\mathrm{d}x + \int_{p_a}^{p_b} \pi_\xi(x)\log\left(\frac{\pi_\xi(x)}{\pi_0(x)}\right)\mathrm{d}x + \int_{p_b}^{\infty} \pi_\xi(x)\log\left(\frac{\pi_\xi(x)}{\pi_0(x)}\right)\mathrm{d}x \\
&\quad = \mathrm{KLD}_{\text{Gumbel}}(\pi_\xi, \pi_0) + \mathrm{KLD}_{\text{Blending}}(\pi_\xi, \pi_0) + \mathrm{KLD}_{\text{Fréchet}}(\pi_\xi, \pi_0).
\end{aligned}$$

A closed-form expression can be found for the Gumbel part of the KLD, where we express the Gumbel distribution function as $G(x) = \exp(-h_2(x))$ and use the same substitution techniques as for the KLD computations with the GEV distribution. We get

$$\begin{aligned}
&\mathrm{KLD}_{\text{Gumbel}}(\pi_\xi, \pi_0) = \\
&\quad -\Gamma_u(-\log p_a; 2) + p_a\log(-\log p_a) - E_i(\log p_a) + \Gamma_u(-\log p_a; C_1+1)e^{-C_2} \\
&\quad - C_1\left(p_a\log(-\log p_a) - E_i(\log p_a)\right) + p_a\left(C_2 + \log\xi + \log\left(\log\left(\frac{\log p_a}{\log p_b}\right)\right)\right) \\
&\quad - p_a\log\left((-\log p_b)^{-\xi} - (-\log p_a)^{-\xi}\right),
\end{aligned}$$

where

$$C_1 = \frac{1}{\xi}\frac{(-\log p_b)^{-\xi} - (-\log p_a)^{-\xi}}{\log\left(\frac{\log p_a}{\log p_b}\right)} \quad \text{and} \quad C_2 = C_1\log(-\log p_a) + \frac{1}{\xi}\left((-\log p_a)^{-\xi} - 1\right).$$

Here, $\Gamma_u(x;\alpha) = \int_x^\infty t^{\alpha-1}e^{-t}\mathrm{d}t$ is the upper incomplete gamma function, and $E_i(x) = \int_{-\infty}^x (e^t/t)\mathrm{d}t$ is the exponential integral, which can be evaluated using the GNU Scientific Library (Galassi et al., 2007; Hankin, 2006, 4). The Fréchet part of the KLD is found in the same way as (S2.1), only using different integration limits. For $0 \leq \xi < 1$ we get

$$\begin{aligned}
\mathrm{KLD}_{\text{Fréchet}}(\pi_\xi, \pi_0) = &-\Gamma_l(-\log p_b; 2) + (\xi+1)\left(-p_b\log(-\log p_b) + E_i(\log p_b) - \gamma\right) \\
&+ \frac{1}{\xi}\left(\Gamma_l(-\log p_b; 1-\xi) - (1-p_b)\right) + \int_{p_b}^1 \exp\left(\frac{1}{\xi}\left(1-(-\log u)^{-\xi}\right)\right)\mathrm{d}u,
\end{aligned}$$

where $\Gamma_l(x;\alpha) = \Gamma(\alpha) - \Gamma_u(x;\alpha)$ is the lower incomplete gamma function. The integral in the expression above must once again be evaluated numerically. We are unable to find an

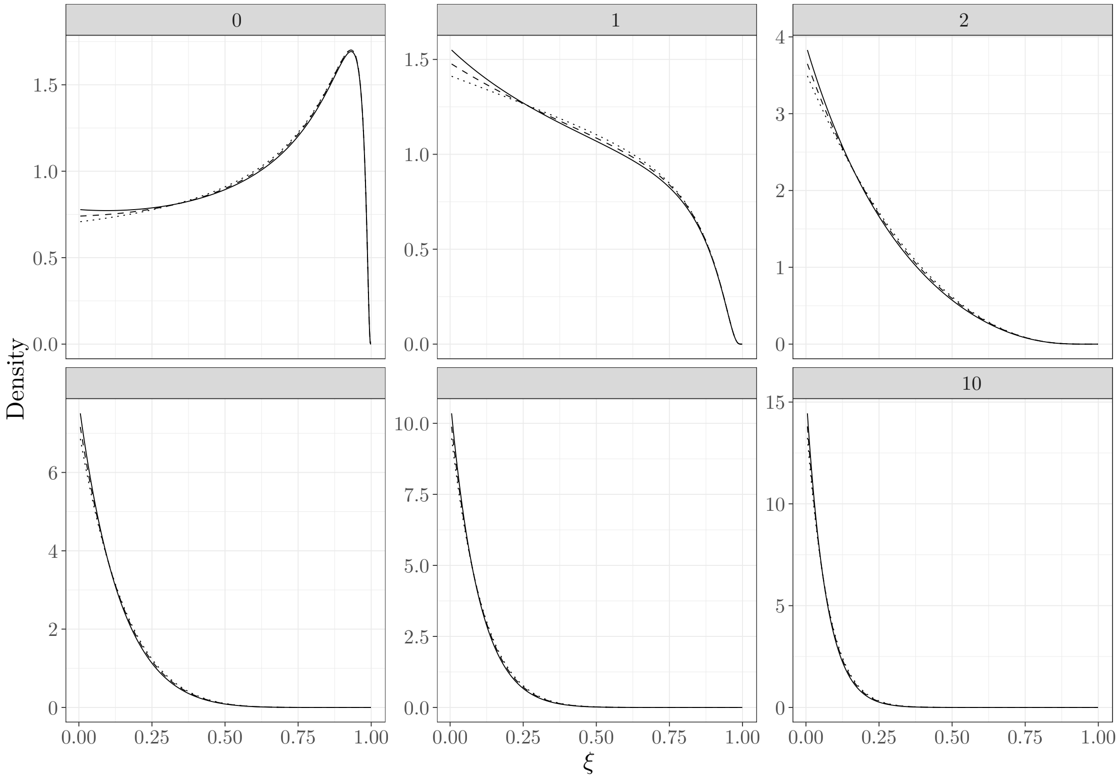

Figure S2.1: PC priors with base models $\xi = 0$ and different penalty parameters λ for the GEV (solid line), bGEV (dashed line) and generalised Pareto distribution (dotted line).

expression for the KLD integral between the p_a quantile and the p_b quantile. Thus, we use numerical integration with $\mu = 0$ and $\sigma = 1$ to compute $\text{KLD}_{\text{Blending}}(\pi_\xi, \pi_0)$. The value of the integral does not depend on the values of μ and σ, but we are unable to compute it without choosing some value for them. Being unable to find an expression for the KLD of the bGEV distribution, we must also use numerical derivation to estimate the derivative of the KLD, needed in (S2.2).

The PC priors for the GEV, bGEV and generalised Pareto distributions are displayed in Figure S2.1 for different values of the penalty parameter λ. For all the chosen values of λ the three distributions are so similar that their effect on a posterior distribution probably will be close to identical. The PC prior for the generalised Pareto distribution exists in closed-form and is already implemented in `R-INLA`, while the other PC priors must be computed numerically and are not implemented in `R-INLA`. Consequently, we choose to use the PC prior of the generalised Pareto distribution for modelling the tail parameter of the bGEV distribution.